O-007

ブログ注目情報と株価変動の相関分析に関する検討

A Study on the Correlation Analysis between Blog Hot Topics and Stock Price Changes

原　慎司[1]　　灘本裕紀[2]　　堀内　匡[1]
Shinji Hara　　Hironori Nadamoto　　Tadashi Horiuchi

1. はじめに

近年，Web上で一般の人々が容易に情報を発信する手段としてblogが注目されている．blogは即時性・リアルタイム性のある新鮮な情報を配信しているため，新たな情報源としても注目されている．このblogを大量に収集し，blogの集合を対象としてさまざまな手法で分析することで，一般の人々の「生の声」を抽出しようという試みであるblogマイニングと呼ばれる新しい研究が始まっている[1]．

本研究では，blogマイニングに注目し，実世界の動向との相関分析の一つとして，株価の変動と相関が高いキーワード群を大量のblogから抽出する手法とそれを利用した株価予測について検討する．具体的には，kizasi.jp[2]というblogサーチエンジンを利用して株銘柄の注目情報を収集し，その株銘柄の関連キーワードと実際の株価の変動から株価の上昇・下降に相関が高いキーワード群を抽出する．さらに，抽出されたキーワード群とblogの注目情報を利用した株価予測をナイーブベイズ法[3]により試みる．

2. 提案手法

本研究では，kizasi.jpより抽出した株銘柄情報を利用して株価変動に関連したキーワードを抽出し，そのキーワードを用いたナイーブベイズ法というテキストマイニング手法により株価の予測を行う．提案する手法の枠組みを図1に示す．

2.1 株価変動キーワードの抽出

kizasi.jpの株チャンネルを利用して，注目されている株銘柄とそれらに関連性の高いキーワードを多数集め，その中の各株銘柄について，実際の株価の変動を調べ，クラスc_1「株価上昇」とクラスc_2「株価下降」に手動で分類する．分類された銘柄情報は，表1のようにデータベースに格納する．

次に，この分類された銘柄情報から，各クラスのキーワードの出現回数を求める．また事前確率を，クラスに分類された銘柄数とその中でキーワードを含む銘柄数の比として算出する．これにより，株価変動と相関の高いキーワードを抽出することが可能になる．抽出したキーワードの例を表2に示す．

2.2 情報収集の自動化

大規模データを用いて実験を行うために，kizasi.jpが公開しているWebAPIを利用して株銘柄情報を自動取得するプログラム，および株価情報をYahoo!ファイナンス[4]より自動収集するプログラムを作成する．

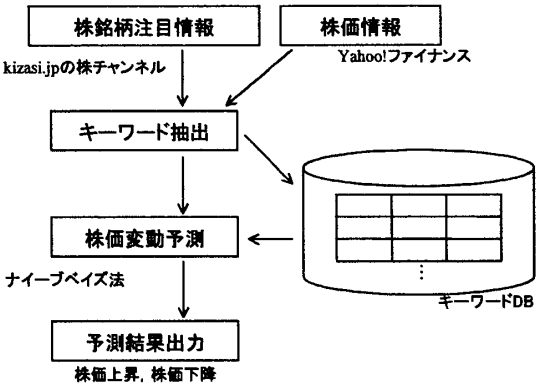

図1: 提案手法の枠組み

表1: 銘柄情報の例

銘柄名	クラス	キーワード
I	株価上昇	上昇, 株価, 買い, スイング, ···
J	株価下降	下方修正, 買い, 空売り, 暴落, ···

表2: キーワード出現頻度の例

キーワード	出現回数		事前確率	
k_i	c_1	c_2	$P(k_i\|c_1)$	$P(k_i\|c_2)$
買い	30	70	0.2	0.5
売り	80	40	0.7	0.3
上方修正	50	10	0.8	0.2

2.2.1 株銘柄情報収集の自動化

kizasi.jpでは関連語等の情報をプログラムで利用するためのkizAPIを公開している．kizAPIは，要求したスパン（24時間/1週間/1ヶ月）でキーワードの関連語をRSS形式で返す．まず，株チャンネルのRSSデータを定期的に取得し，注目されている株銘柄名を取得する．次に，その株銘柄の関連語のRSSデータをkizAPIを用いて取得する．また，一度株チャンネルにランクインした株銘柄の関連語は，継続的に取得するようにした．

2.2.2 株価情報取得の自動化

現在，日本国内の株銘柄情報を取得するためのフリーのWebサービスはない．そこで，Yahoo!ファイナンスの株銘柄情報のページより，HTMLパーサ[5]を利用して株価を自動取得した．

[1]松江工業高等専門学校，Matsue College of Technology
[2]京都大学大学院，Kyoto University

3. 実験

提案手法の有効性を検証するために，2.2節で示した自動化により取得した大規模データを用いた予測実験および株銘柄情報としての関連語と実際の株価変動との相関分析を行った．

3.1 データの収集

より多くの株銘柄情報を利用した分類実験を行った．まず，情報収集の自動化により取得した関連キーワードを銘柄情報取得日の前日の株価 ($p0$) と翌日の株価 ($p1$) を用いて，次式 (1)〜(5) で分類し，表1のようにデータベースに格納した．また，価格にはその日の5日平均値を用いた．

$$p' = \frac{p1 - p0}{p0} \quad (1)$$

$$p' \geq 0.02 \Rightarrow \text{「上昇」} \quad (2)$$

$$p' < 0.02 \Rightarrow \text{「その他1」} \quad (3)$$

$$p' \leq -0.02 \Rightarrow \text{「下降」} \quad (4)$$

$$p' > -0.02 \Rightarrow \text{「その他2」} \quad (5)$$

分類の組み合わせとしては，次に示す2種類の分類を行った．1つ目は，式 (2), (3) を用いたクラス「上昇」とクラス「その他1」への分類（以下，分類Aと呼ぶ）である．式からわかるように，クラス「その他1」はクラス「下降」に属する情報も含んだものである．2つ目は，式 (4), (5) を用いたクラス「下降」とクラス「その他2」への分類（以下，分類Bと呼ぶ）であり，クラス「その他2」はクラス「上昇」に属する情報も含んだものである．次に，この分類結果をもとに，表2のようにキーワードの出現頻度を求めた．これらにより，作成されたデータの概要を以下に示す．

- 収集期間：2007/08〜2007/10の2ヶ月間
- 総銘柄情報数：5922件
  （「上昇」：1469，「その他1」：4453）
  （「下降」：1421，「その他2」：4501）
- 総登場キーワード数：25297個
- 全体の98%は出現率が1%以下

これらの収集した株銘柄情報は，分類Aと分類Bともにキーワードの出現数およびクラス出現数に大きな偏りがある．そこで，分類属性として利用するキーワードおよび実験に利用する事例の選択を行った．

3.1.1 キーワードの選択

分類属性とするキーワードの選択には，閾値での打ち切りと情報エントロピーを用いた重要語の抽出を行った．

まず，閾値として全体の1%以上として60件以上を設定した．この閾値によるデータ選択後のキーワード数は472個であった．次に，情報エントロピーを用いた重要語の選択を行った．ここで c^{m_w} は，キーワード w を含む銘柄情報列であり，$c^{\neg m_w}$ は，キーワード w を含まない銘柄情報列である．このとき，c^{m_w} および $c^{\neg m_w}$ の情報エントロピーをそれぞれ $I(c^{m_w})$，$I(c^{\neg m_w})$ とするとき，これらの和を全銘柄データ数 m で割った値をそのキーワードの重要度とする．このとき，c^{m_w} は式 (7) を用いて計算される．この式において，$H(z)$ は式 (8) で与えられるエントロピー関数である．m_w^u は w を含む銘柄情報のうちクラス「上昇」に分類された銘柄数であり，m_w はキーワード w を含む銘柄数である．

表3: 選択されたキーワード

keyword 「上昇」「その他1」	keyword 「下降」「その他2」
ストップ高	S安
引け	下方修正
銘柄	終了
cm	富士
売り	米
上昇	発売
新興	通
1株	ストップ安
フルスピード	日本
買われた	下げ
マザーズ	新規
続伸	パソコン
S高	初値
登録	決済完了
JQ	公募
アイフリーク	後場
OJ	下落
アイル	2007年
高値	売り
寄り	車
買い	行きました
資金	高級
ちり	KDDI
マネーパートナーズ	底
強い	平均
持ち越し	手数料
マネパ	続落
ヘラクレス	審査
含ま	買戻し
線	グッドウィル

$$E(w) = \frac{1}{m}\{I(c^{m_w}) + I(c^{\neg m_w})\} \quad (6)$$

$$I(c^{m_w}) = m_w H\left(\frac{m_w^u}{m_w}\right) \quad (7)$$

$$H(z) = -z log_2 z - (1-z) log_2 (1-z) \quad (8)$$

この情報エントロピーは，事象の乱雑さを表す指標である．この実験においては，両クラスに等確率で出現する場合に最大となり，出現に偏りがあるほど小さくなる．つまり，式 (6) の値の小さいものを選択することにより，「上昇」もしくは「その他1」に強く寄与するキーワードを選択することができる．また，分類Bについても同様に，「下降」もしくは「その他2」に強く寄与するキーワードを選択することができる．

この方法により選択された30個のキーワードを表3に示す．

3.1.2 事例の選択

収集したデータは，クラス間で銘柄数に大きな偏りがある．また，選択した30個のキーワードがまったく出現しない銘柄も多く含まれている．そこで，30個のキーワード群と各銘柄情報との共通要素数に対して閾値を設定して事例を選択した．この共通要素数とは，選択した

表 4: 各閾値における銘柄情報の比較

閾値	総銘柄数	上昇	その他 1	上昇/総銘柄数
0	5922	1469	4453	0.248
1	3348	1032	2316	0.308
2	2155	761	1394	0.353
3	1406	569	837	0.405
4	909	434	475	0.477
5	586	331	255	0.565

表 5: 閾値 0 での結果：分類 A

		分類されたクラス		合計
		「上昇」	「その他 1」	
実際のクラス	「上昇」	463	1006	1469
	「その他 1」	489	3964	4453
合計		952	4970	5922

精度 (Accuracy): 0.748

表 6: 閾値 0 での結果：分類 B

		分類されたクラス		合計
		「下降」	「その他 2」	
実際のクラス	「下降」	294	1127	1421
	「その他 2」	290	4211	4501
合計		584	5338	5922

精度 (Accuracy): 0.761

30 個のキーワードをベクトル $\boldsymbol{w}$，銘柄 S_i に含まれるキーワードベクトルを $\boldsymbol{s}_i$ で表すと，共通要素数はこれらの積集合 $\boldsymbol{w} \cap \boldsymbol{s}_i$ の要素数である．これを全銘柄に対して計算し，設定した閾値以上である銘柄のみを選択した．これにより，選択したキーワード群と共通のキーワードを多く含む銘柄を残すことで，より相関の高い銘柄情報のみ残ることになる．この閾値は 0〜5 までの 6 パターンを設定した．また，閾値 0 とは全銘柄情報を利用することと同義である．

分類 A（「上昇」と「その他 1」への分類）について，各閾値における銘柄情報数の概要を表 4 に示す．この表より，バランスのとれた株銘柄情報集合を構築することが可能になることがわかる．

3.1.3 実験結果

分類 A（「上昇」と「その他 1」への分類）および分類 B（「下降」と「その他 2」への分類）について，閾値 0 でのそれぞれの実験結果を表 5，6 に示す．また，分類 A について，各閾値ごとの評価結果を比較した様子を図 2 に示す．

図 2 より，精度 (accuracy) は閾値が大きくなるに従って低下していることがわかる．一方，完全性の指標である再現率 (recall) および正確性の指標である適合率 (precision) は上昇している．また，表 5，6 より，「上昇」と「その他 1」への分類 A，「下降」と「その他 2」への分類 B について，精度を求めると，それぞれ 0.748 と 0.761 であり，同程度の結果が得られたことがわかる．

3.1.4 考察

3.1.2 節で示したように，閾値が大きくなるほどクラス間の株銘柄情報数の偏りは小さくなる．つまり，分類

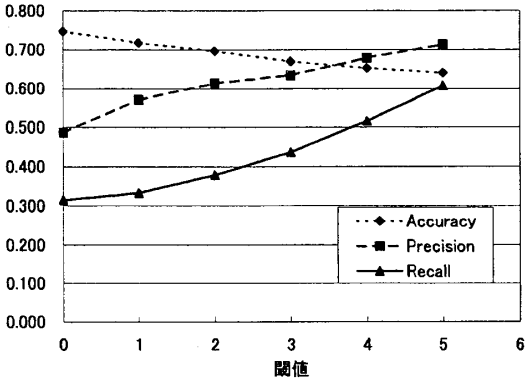

図 2: 分類結果の比較

対象となる株銘柄情報集合のエントロピーが大きくなり，予測は難しくなる．実験結果において，図 2 のように閾値の増加に対して精度が減少したのはこのためであると考えられる．

つぎに，再現率および適合率について考える．本実験において，閾値を大きくするに従って，再現率および適合率は上昇している．つまり，分類 A の実験において，閾値を大きくしていくにつれ，正しくクラス「上昇」に分類する正分類が多かったといえる．従って，実験において用いたキーワード群はクラス「上昇」と高い相関があったと考えられる．

4. おわりに

本研究では，blog 集合から抽出された株銘柄情報および株価情報より，注目キーワード群と実際の株価変動の相関を分析した．今回の実験結果において，選択されたキーワード群は株価上昇のクラスと相関が高いと考えられるものであったが，今後さらに株銘柄情報を継続して収集し，分類実験を行う必要がある．また，キーワードの選択方法や閾値の設定方法，事例選択法などの正当性についても検討しなければならない．

参考文献

[1] 奥村学, blog マイニング－インターネット上のトレンド，意見分析を目指して－，人工知能学会誌, Vol.21, No.4, pp.424 - 429 (2006)

[2] 株式会社きざしカンパニー, “kizasi.jp” http://kizasi.jp/

[3] I.H. Witten and E. Frank, *DATA MINING: Practical Machine Learning Tools and Techniques*, pp.365 - 483, Morgan Kaufmann (2005)

[4] ヤフー株式会社, “Yahoo!ファイナンス” http://quote.yahoo.co.jp/

[5] HTML Parser, http://htmlparser.sourceforge.net/

[6] 元田浩ほか，IT Text データマイニングの基礎，pp.123-166，オーム社 (2006)