F-014

ブロガーの注目情報を用いた株価変動予測の試み
A Study on Stock Price Prediction using Blogger's Trend Information

灘本　裕紀 [1]　　**堀内　匡** [1]
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（Weblog）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，この新しい 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変動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．そして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の予測をナイーブベイズ法 [3] と呼ばれる手法により行う．

2. blog マイニング

blog を対象としたマイニングでは，blog 記事を対象とした分類・意見抽出などのテキストマイニングのアプローチのみではなく，リンク構造からのコミュニティー抽出や時間情報からのトレンド分析などの様々なマイニングが可能である．本研究では，blog から株銘柄の注目情報を抽出するためのトレンド分析とテキストマイニング手法の一つであるナイーブベイズ法を用いて株価変動予測を行う．このトレンド分析には kizasi.jp という既存の blog サーチエンジンを利用する．

3. 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．図 1 に提案する手法の枠組みを示す．

3.1 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．

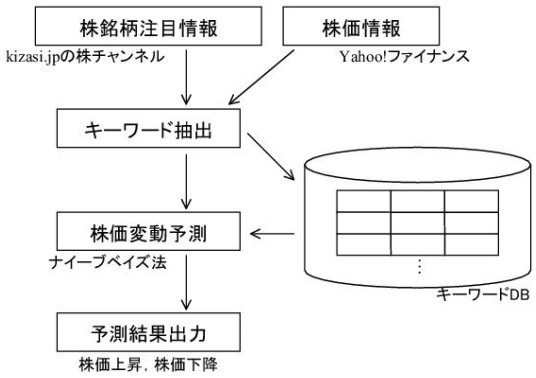

図 1: 提案手法の枠組み

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

3.2 ナイーブベイズ法による予測

ある銘柄 S を「株価上昇」と「株価下降」に分類する場合を考える．この銘柄に含まれるキーワード群を $\boldsymbol{s} = (s_1, \cdots, s_n)$ とすると，事後確率 $P(c|\boldsymbol{s})$ を最大とするクラス $\hat{c}$ を求めることで，分類の誤りを最小とすることができる．この事後確率 $P(c|\boldsymbol{s})$ は，ベイズの定理を適用し，式 (1) のように求めることができる．

$$P(c|\boldsymbol{s}) = \frac{P(c)P(\boldsymbol{s}|c)}{P(\boldsymbol{s})} \tag{1}$$

また，すべてのクラスで $P(\boldsymbol{s})$ が一定であることを考慮すると，$\hat{c}$ は式 (2) のように「そのクラスの出現確率」$P(c)$ と「クラス別のキーワード群の出現確率」$P(\boldsymbol{s}|c)$ の積を最大とするクラスとして求めることができる．

$$\hat{c} \quad = \quad \arg\max_c P(c|\boldsymbol{s})$$

[1] 松江工業高等専門学校，Matsue National College of Technology

$$= \arg\max_c \frac{P(c)P(\boldsymbol{s}|c)}{P(\boldsymbol{s})}$$
$$= \arg\max_c P(c)P(\boldsymbol{s}|c) \quad (2)$$

このとき，各キーワードが互いに独立と仮定すると，式 (2) の右辺の確率は次のようになる．

$$P(\boldsymbol{s}|c) = P(s_1, \cdots, s_n|c) \approx \prod_{i=1}^{n} P(s_i|c)$$

これは，前節で示したように算出可能である．したがって式 (3) を最大化するクラス $\hat{c}$ を求めればよい．

$$\hat{c} = \arg\max_c P(c) \prod_{i=1}^{n} P(s_i|c) \quad (3)$$

この方法による予測は，より正確な確率を学習していくことにより高精度になっていく．

この手法は，スパムメールフィルタに利用されているベイジアンフィルタ [4] と同様である．ベイジアンフィルタにおいても，学習が進むことで高精度なスパム検出を行うことが可能になっている．

4. 予備実験

この手法の有効性を確認するために，kizasi.jp より実際に株銘柄情報を取得し，予備実験を行った．

4.1 実験方法

実験のために kizasi.jp より 1 週間おきに 1 回 30 件ずつ合計 3 日分の株銘柄情報（30 件× 3 日=90 件分）を取得した．そして，実際の株価変動チャートを見ることで，これらの株銘柄情報を「株価上昇」と「その他」という 2 つのクラスに分類した．また，登場する各キーワードについてクラスごとの出現数を求めた．このとき作成した表の一部を表 3 および表 4 に示す．

次に，求めた出現回数から特徴的なキーワードを 29 個選択してナイーブベイズ法による分類実験を行った．分類実験には，マイニングツール Weka[5] を利用した．

4.2 実験結果

実験結果は，表 5 のようになった．この結果より，正しく分類された銘柄は 21+53=74 件（82.2%）である．これは予備実験としては良好な結果であり，この方式の有効性は確認できた．また今回の実験では，クラス「その他」の銘柄をクラス「株価上昇」に分類する誤りはなかった．

表 3: 実際の銘柄情報の例

銘柄	クラス	キーワード			
O1	その他	株価	上昇率	仕手化	東 1
O2	その他	マザーズ	株	銘柄	経常
O3	その他	東 1	2 月	売り	上位
U1	株価上昇	下げ	買い	標的	場提供
U2	株価上昇	PER	東証 1 部	空売り	狙い目
U3	株価上昇	東証 1 部	株	上方修正	発表

表 4: 実際の出現頻度の例

キーワード	株価上昇	その他
初値	0	8
公募価格	0	8
終値	0	7
トレード	5	0
続伸	5	0
小売業	0	5
野村	0	5
公開価格	0	5
東証 1 部市場	4	0
業種	4	0

表 5: 実験結果

		分類されたクラス		
		株価上昇	その他	合計
実際のクラス	株価上昇	21	16	37
	その他	0	53	53
合計		21	69	90

5. おわりに

本研究では，blog の集合から株銘柄に対する注目情報を抽出し，ナイーブベイズ法というテキストマイニング手法を用いて株価変動予測を行う手法を検討した．予備実験として実データを利用した分類テストを行い，有効な結果を得た．

kizasi.jp では，RSS という情報配信フォーマットを利用した WebAPI を公開している．この RSS データは，XML 形式で記述されているためにプログラム上での解析が容易であり，これを利用することでキーワード情報や出現頻度などを定期的に自動取得することができる．現在，この RSS を取得し解析して銘柄情報を取得するプログラムを作成し，データ収集を始めている．

また，キーワード抽出でのクラス分類を自動化するために，Yahoo!ファイナンス [6] からの株価情報自動取得を検討している．

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] 竹村彰通ほか, 統計科学のフロンティア 10 言語と心理の統計, pp.59 - 128, 岩波書店 (2003)

[4] 山井成良, 桝田秀夫, “spam メールの現状と対策の動向”, 情報処理学会誌, Vol.46, No.7, pp.739 - 772 (2005)

[5] Ian H. Witten, Eibe Frank, *DATA MINING: Practical Machine Learning Tools and Techniques, Second Edition*, pp.365 - 483, MORGAN KAUFMANN (2005)

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