ブログ上の単語出現頻度と株価の関係性の定量評価

The quantitative evaluations of relationships between word appearances on blogs and stock prices

渡邊隼史[1*]　セーヨーサンティ[1] 山田健太[1,2] 内山 幸樹[1]
Hayafumi Watanabe[1] Saeyor Santi[1] Kenta Yamada[1,2] Koki Uchiyama[1]

[1] 株式会社ホットリンク [2] 早稲田大学高等研究所
[1] Hottolink, Inc. [2] Waseda Institute for Advanced study

Abstract: To clarify connections between social media data and financial market data, we studied the quantitative evaluations of relations between time series of word appearances on Japanese blogs and those of stock prices. In particular, we proposed the method of comparison of some correlation indices such as the Spearman's rank correlation coefficient, based on the man-made related stock information. We found that the Spearman's rank correlation coefficients over time series of 562 keywords can hardly pick the correct combinations of related stocks out the pool of more than 3,000 stocks on the Tokyo Stock Exchange. However, we show that the composite correlation indicators, which reflect multiple features of the time series, can pick the correct stocks up to a certain level of statically significant.

1　はじめに

近年，ソーシャルメディアデータは，一般の消費者や社会の声を代表するものとして，マーケティングや選挙，報道など様々な分野で利用されはじめている．ソーシャルメディアを利用することで，金融関係でも，ニュースに対する一般社会の反応やその残存期間，長期的な人々の関心の変化やブームの動向，ネット炎上の早期の検知，時代のムードなどこれまで金融関係で使われていた情報源では得られなかった有用な情報を得られる可能性がある [1, 2]．そこで，本研究では，ソーシャルメディアと金融に関する研究の第一歩として，ブログ上の単語の出現頻度時系列と株価等の金融時系列の関係性とその評価の手法について研究を行う．特に，詳細な解析をおこなう前処理として，多くのキーワードと株価の組み合わせから，書き込み頻度時系列と株価データ（取引量データも含む）のみを利用して，キーワードとそのキーワードと関係性がある銘柄の組み合わせを絞り込む方法の提案を行う．

本稿では，まず，研究の対象のデータ，今回注目する関係性の定義とその関係性の評価等の解析方法について述べる．次に，データ解析の結果として，統計的に有意にブログデータと株価データが関係していること，提案手法を用いることで様々な相関指標を系統的に比較できることを示す．最後に，むすびとして，本論文のまとめと今後の展望を述べる．

*連絡先：株式会社ホットリンク
〒 102-0081　東京都千代田区四番町 6 番 東急番町ビル
E-mail: h.watanabe@hottolink.co.jp

2　提案方法

2.1　研究対象のデータ

本研究では，ソーシャルメディアと金融市場データの研究の第一歩として，それらの時系列の情報として最も基本的な情報である，ブログキーワード時系列と上場企業の株価と取引量のデータを扱う．ここでは，「ブログキーワード時系列（キーワード時系列）」とは，ある日にちに書き込まれる特定のキーワードを含む日本中のブログの件数の時系列である．ブログキーワード時系列はホットリンク社の口コミ係長ＡＰＩサービスから取得した．口コミ係長ＡＰＩサービスを用いることで，任意のキーワードの，2006 年 11 月 1 日から現在（本研究では 2012 年 9 月までのデータを用いた）までのキーワード時系列，ポジティブな言葉を含む記事に限定したキーワード時系列，また，ネガティブな言葉を含む記事に限定したキーワード時系列を取得できる．また，サービスに含まれるスパムフィルタを用いることで，広告等のスパムブログを除去することができる（本研究では，スパムフィルタにより，スパムブログ除去後のキーワード時系列について解析を行っている）．同期間の株価データについては，株

価データ倉庫 (http://www.geocities.co.jp/WallStreet-Stock/9256/data.html) から取得した．本研究では，株価時系列は市場全体の動きの効果を除去するため（キーワードの効果を明確にするため），株価を日経平均にほぼ対応する時系列（具体的には，観測期間中のすべての日にちで値をもつ銘柄群について，各銘柄の観測期間の平均株価で除算した「規格化した銘柄の株価」を計算し，その「規格化した銘柄の株価」の銘柄に関する中央値を各日にちごとにとった時系列）で除算したものとする．また，口コミ時系列は，ブロガー数の増減の効果を除去するため，キーワード時系列を（平均１にスケール化した）観測したブログの総数で除した時系列を用いる．これ以降，これらの規格化した時系列を単に，株価時系列，口コミ時系列と呼ぶ．株式の取引数は市場全体の効果はあまりみられなかったので，このような規格化をおこなわず生の時系列を用いた．

2.2 関連性と関連性の指標の評価

株式情報とブログ件数の時系列の関係性 本研究では，ソーシャルメディア時系列と株式情報時系列の間になんらかの関係性がある蓋然性の高さを定量的に評価する方法を開発することを目的としている．特に，本研究でいう関係性とは，２つの時系列を発生させる機構になんらかの因果関係が存在することと考えることにする．例えば,「インフルエンザが大流行し，マスク企業（繊維メーカー）の売上が増すことを予測した投資家が投資し株価が上昇すること」と「インフルエンザ流行によってインフルエンザのブログ上での書き込み数が増加すること」は２つの増加にインフルエンザの大流行という共通の原因があるので,「マスク企業（繊維メーカー）の株価」と「インフルエンザのキーワード時系列」の２つの時系列に関係性があると考える．一方，インフルエンザの流行は毎年２月ごろにピークの１年間の周期がある，それに伴い「インフルエンザ」の書き込み数も２月にピークがある年周期をもっている．また,「マックスバリュ北海道」のように２月末に株主優待の権利確定日がある銘柄も２月にピークを向かえる周期をもつことも多い．これらのキーワード時系列と銘柄の株価は時系列として２月にピークあるという点で時系列の形状は類似しているが，それらの間に本質的な因果関係は存在しない．したがって，この場合，「インフルエンザ」と「マックスバリュ北海道」は関係性がないと考える．本研究では，時系列だけの情報を使って，ブログのキーワード時系列と上記に意味で関係性がある株価式銘柄を機械的に抽出する方法を考える．より具体的には，キーワード時系列と株価時系列のひとつのペアを選んだとき（相関係数のような）関係性の有無や大きさを評価する指数を付与することを考える．

関係性指標の評価方法 本研究では，いくつかの関係性指標を導入し，キーワード時系列と株価の関係性の定量的な評価を試みる．そこで，まず，導入する関係性指標間の性質を比較するため，関係性指標を系統的に評価する方法を導入する．

キーワードと企業の関連性は，企業の事業内容を把握し一つ一つ調べるべきである（例えば，Web ページなどを参照しながら）．しかしながら，無数あるキーワードと数千ある銘柄との組み合わせのすべてについて事業内容とキーワード関係性を調べることは非常に困難である．そこで，私たちはその代わりとして，Web 上で公開されている「関連銘柄情報データ」を利用した．本研究では，この関連銘柄のデータとして，Nifty ファイナンスの「@nifty ファイナンス オリジナルテーマ分類」を用いた．ここで,「関連銘柄情報データ」とは，投資のために，ある投資のテーマのキーワードとテーマごとにそこから連想される株式銘柄の組（関連銘柄リスト）を示した人手で作った情報である．この情報を用いることで，時流のテーマと（人間が業務内容上関係あると認識している）それに関連する投資銘柄を知ることができる．たとえば，テーマとして「インフルエンザ」を与えられ，その関連銘柄として，繊維メーカーの「ダイワボウホールディングス」や「シキボウ」が与えられる．Nifty ファイナンスの関連銘柄情報データには 593 テーマが存在した (2012 年 9 月時点)．

593 テーマの口コミ件数と全上場企業のすべての組み合わせペアついて，ある関連性指標が付与されているとする．そのとき,「指標の値が敷居値以上のペアのうち関連銘柄リストにも含まれているペアの割合」を関連性指標の関係性の抽出能力の評価値とする．具体的には，敷居値を r として，

$$q(r) = \frac{\text{指数が } r \text{ 以上かつ関連銘柄リストに含まれる銘柄数}}{\text{指数が } r \text{ 以上の全銘柄数}} \tag{1}$$

を用いる．また，この量とランダム抽出と相違は，基準の確率 p_0 とサンプル数に応じた二項検定を適用することで p 値として測定できる．基準の確率 p_0 は

$$p_0 = \frac{\text{関連銘柄に含まれる銘柄数とテーマのペア数}}{\text{テーマ数と銘柄の全ペアの組み合わせの数}} \tag{2}$$

で求められる．この量は，時系列情報のみから得られた関連性指標の値がどの程度人間が付与した関連性と近いか，ということに対応している．この量が大きいほど人間の目で付与した関連銘柄に近いということを意味する．

2.3 関連性の指標

以下の時系列の関連性の指標を比較した．

- 時系列そのものの相関係数
- 箱平均の差分の相関係数
- 確率的にまれな事象（異常値等）の発生の同期性
- 上記の指標を組み合わせた指標

以下，上記の関連性の指数のそれぞれについて記述する．

相関係数　相関係数は，もっとも基本的な関係性（時系列類似性）の評価手法である．本研究では，相関係数のうち，スピアソンの順位相関係数を用いる．ここで，時系列 $x(t)$ と $y(t)$ $(t = 1, 2, \cdots, T)$ のスピアマンの順位相関係数は，

$$cor[x, y] = \frac{< (R(x) - < R(x) >) \cdot (R(y) - < R(y) >) >}{\sqrt{< (R(x) - < R(x) >)^2 >< (R(y) - < R(y) >)^2 >}} \tag{3}$$

で与えられる．ここで $R(\cdot)$ は時系列を順位に変換する関数である．たとえば，$T = 4$ の場合，

$$x(1) = 5, x(2) = 7, x(3) = 6, x(4) = 2$$

なら，

$$R(x(1)) = 2, R(x(2)) = 4, R(x(3)) = 3, R(x(4)) = 1$$

となる．また，$< \cdot >$ は時間平均である．なお，通常よく使われるピアソンの積率相関係数を用いない理由は，我々の扱う時系列では，ピアソン積率相関が利用できる前提条件－線形相関かつ時系列の従う分布が正規分布－を必ずしも満たさないと予想されるからである．ところで，相関係数には，無関係な時系列（たとえば，２つの独立のランダムウォーク）でも相関係数が大きくなる現象（偽相関）が現れてしまう悪い性質が知られている．

ボックス平均の差分の相関係数　非定常時系列に現れる偽相関の問題を回避するため一般の時系列解析では，時系列の差分をとって定常化してから相関をとることが多い．しかしながら，ブログ時系列の差分は，ノイズ量が大きく，差分の相関係数はほとんど０になってしまう．それらを回避するため，以下のボックス平均の差分の相関係数 $cor[\delta x^{(L)}, \delta y^{(L)}]$ を用いた．ここで，$x^{(L)}$ は時系列 $x(t)$ に対する，ボックスサイズ L のボックス平均

$$x^{(L)}(t') = \sum_{j=t'L+1}^{(t'+1)L} x(j)/L \quad (t' = 1, 2, \cdots, (T-1)/L) \tag{4}$$

であり，$\delta x^{(L)}$ はその差分，$\delta x^{(L)}(t') = x^{(L)}(t'+1) - x^{(L)}(t)$　$(t' = 1, 2, \cdots, (T-1)/L - 1)$ である．

ボックスサイズ L は小さすぎるとゆらぎに影響され，相関係数がほとんど 0 になってしまい，大きすぎるとサンプル数が少なくなる等で推定誤差が大きくなるという問題がある．そこで，本研究では，上記のバランスをとるため，以下（平均相関）を関係性指標として採用した．

$$\bar{cor}[\delta x, \delta y; \bar{L}] = \sum_{i=1}^{\bar{L}} cor[\delta x^{(i)}, \delta y^{(i)}]/L \tag{5}$$

ここで，$\bar{L}$ は，時系列 $\delta x^{(L)}$ を単位根検定をしたときに，L が単位根過程であることが有意水準 5 パーセントで棄却されないもっとも小さい L である．

異常値の同期性　差分相関係数は時間に関してボックス平均をとるため，短期の時間変動の情報が失われてしまう傾向がある．そこで，１日で大きく時系列が変化した日の同期性をしらべることで，短期の時間変動に関する関連性をとらえることを試みる．ここでは，そのような目的の関係性の指数をランダム拡散モデルを用いて導入した．ランダム拡散モデルはブログの時系列のゆらぎの確率モデルであり，これを用いることでニュース等の外的な要因がない連続する２日間の書き込み数の差分の確率分布 $p(x(t+1), x(t))$ を計算できる [3, 4]. この計算された確率分布の対数 $\log(p(x(t+1), x(t)))$ と株価の時系列の商 $y(t+1)/y(t)$ の内積を異常値の同期性の指標とする．生の時系列 $x(t)$ でなく，ジャンプ確率 $p(t)$ を用いることで，確率的なレアな事象（異常値）が強調され，この内積はブログ上の異常値の共通部分だけとりだすフィルタとして機能する．具体的に以下の手順で計算した．

1. $x(t)$ のジャンプ確率の対数を計算する： $\omega(t) = -\log(p(x(t+1), x(t)))$.
2. $y(t)$ の上昇率を計算する： $v(t) = y(t+1)/y(t)$.
3. $\omega(t)$ と $v(t)$ の内積（相関係数に比例）を計算する：$q = \sum_{t=1}^{N} \omega(t) \cdot v(t)$.
4. 無相関のときに $r(t)$ の分布をブートスラップで計算する (以下の r を M 回サンプルする ; ここで $M = 10^4$):
   $r^{(j)} = \sum_{t=1}^{N} random[\omega(t)] \cdot v(t)$ $j = 1, 2, \cdots, M$.
   ここで，$random(\cdot)$ は時系列のランダムシャッフルする関数.
5. 上記のブートストラップ分布から無相関のときに内積が q 以上になる確率を計算．$u = \#\{r : r > u\}/M$．ここで，$\#\{\cdot\}$ は集合 $\cdot$ の個数である．
6. u を関係性の指数とする．

複合指標　上記のいくつかの指標を組み合わせた複合指標を導入する．$x(t)$ をブログの件数，$y(t)$ を日次の株価，$z(t)$ を日次の取引量としたとき，以下の２つの複合指標を定義した．

$$r^{(1)\prime} = c\bar{o}r[\delta x,\delta y;\bar{L}] \cdot c\bar{o}r[\delta x,\delta z;\bar{L}], \tag{6}$$

$$r^{(2)\prime} = r^{(1)\prime} \cdot \log(u). \tag{7}$$

ただし，$r^{(1)\prime}$ では，$c\bar{o}r[\delta x,\delta y;\bar{L}] < 0$ かつ $c\bar{o}r[\delta x,\delta z;\bar{L}] < 0$ の場合 $r^{(1)\prime} = -\infty$ とする．$r^{(1)\prime}$ は取引量と価格が同時に上昇する場合相関が高くなるため，関心の増加が原因で時系列が変動しているものを抽出してると考えられる．$r^{(2)\prime}$ の複合指標は中長期の相関に関する指標と短期の相関の指標の組み合わせているため，長期的にも短期的にも相関が高い組み合わせをほど評価値が大きくなる．

2.4　結果

上にあげた関係性の指数を比較したものを図１に示す．図より，単純な順位相関では，多少のランダムより比率が高いものの，関係性の抽出の能力はランダムに選択したとは大きくは変わりがないことがわかる（一段目）．この識別力の低さは上述の偽相関を大量に抽出していることが原因と考えられる．一方，２段目に示したボックス差分の相関の平均では，関係性の抽出能力は順位相関とそこまでは変わらないもの，右図の p 値は１段目の単純な順位相関と大きく異なり（２９桁程度異なっている），有意に相関を抽出できていることが確認できる．３段目の異常値の相関も２段目に示したボックス差分の相関の平均と比較することでボックス差分と同じ程度の判別性能ということがわかる．４段目の株価と取引量の平均ボックス差分相関の複合指標は，これまでの例と異なり，左図に示した識別精度が５倍から１０倍程度上がっていることがわかる．また，p 値は２段目の株価だけに比べてさらに 31 桁程度小さくなっていることがわかる．５段目は４段目の複合指標に異常値の相関も合わせた指標である．図より判別精度が４段目の株価と取引量の平均ボックス差分相関の複合指標よりあがっていることがわかる．

これらの結果に加え，定性的には，複合指標を用いることで，たとえば，テーマ「インフルエンザ」では，指数上位から，ダイワボーホールディングス，シキボウ，倉庫精練，重松製作所となり，倉庫精練（繊維メーカー）や重松製作所（マスクメーカー）という関連銘柄データにはあげられていないがインフルエンザと関連性がある銘柄を抽出することができた．

3　むすび

本研究では，ソーシャルメディアデータと金融情報の関係を研究する第一段階として，ブログ上の単語件数の時系列と株価等の情報の時系列のみからわかる関係性を定量的に評価する方法を提案し，また，その指標を評価する方法を導入した．そのことにより，ソーシャルメディアのキーワードの書き込み頻度時系列を用いることで，そのキーワードと関係ある銘柄を統計的に有意に抽出できる可能性があることを示した．また，この方法により，一部のキーワードについては，人手でつけた「関連銘柄データ」には存在しないが，関係性のある銘柄（業務内容がキーワードの事項と関連性の強い銘柄）を抽出ことができた．

しかしながら，本研究は不十分なことも多く，精度上実用レベルに達していない．具体的には，導入した定量指標で関係のある銘柄を抽出すると，真に関係性があるものを抽出できるものの，同時に無関係なものも少なくなく抽出してしまう．そのため，精度を高め，このような無関係ものを抽出をできる限り抑える必要がある．また，本研究では，大まかに全銘柄や全テーマについて全体として統計特性を調べた．しかし，現実には，テーマや銘柄によって，指標により向き、不向きがあると考えられる．そのようにテーマや銘柄ごとのより詳細な指標の有効性を調べる必要もある．さらに，本研究では，全観測期間（５年間）全体での関係性を評価した．しかしながら，現実には特定の時期だけ関係性があるような事例も存在すると考えられる．そのような部分時系列を考慮した関係性の評価についても研究を行う必要がある．

参考文献

[1] Bollen J., Huina M., and Xiaojun Z.: Twitter mood predicts the stock market *Journal of Computational Science* Vol. 2 No. 1, pp. 1-8 (2011).

[2] Preis T., Moat H. S. and Stanley H. E.: Quantifying Trading Behavior in Financial Markets Using Google Trends, *Sci. Rep.*, Vol. 3, pp. 10.1038/srep01684 (2013)

[3] Sano Y., Yamada K., Watanabe H., Takayasu H. and Takayasu M.: Empirical analysis of collective human behavior for extraordinary events in the blogosphere, *Phys. Rev. E*, Vol. 87, No. 1, pp. 012805 (2013)

[4] Watanabe H., Sano Y., Takayasu H. and Takayasu M.: to be submitted

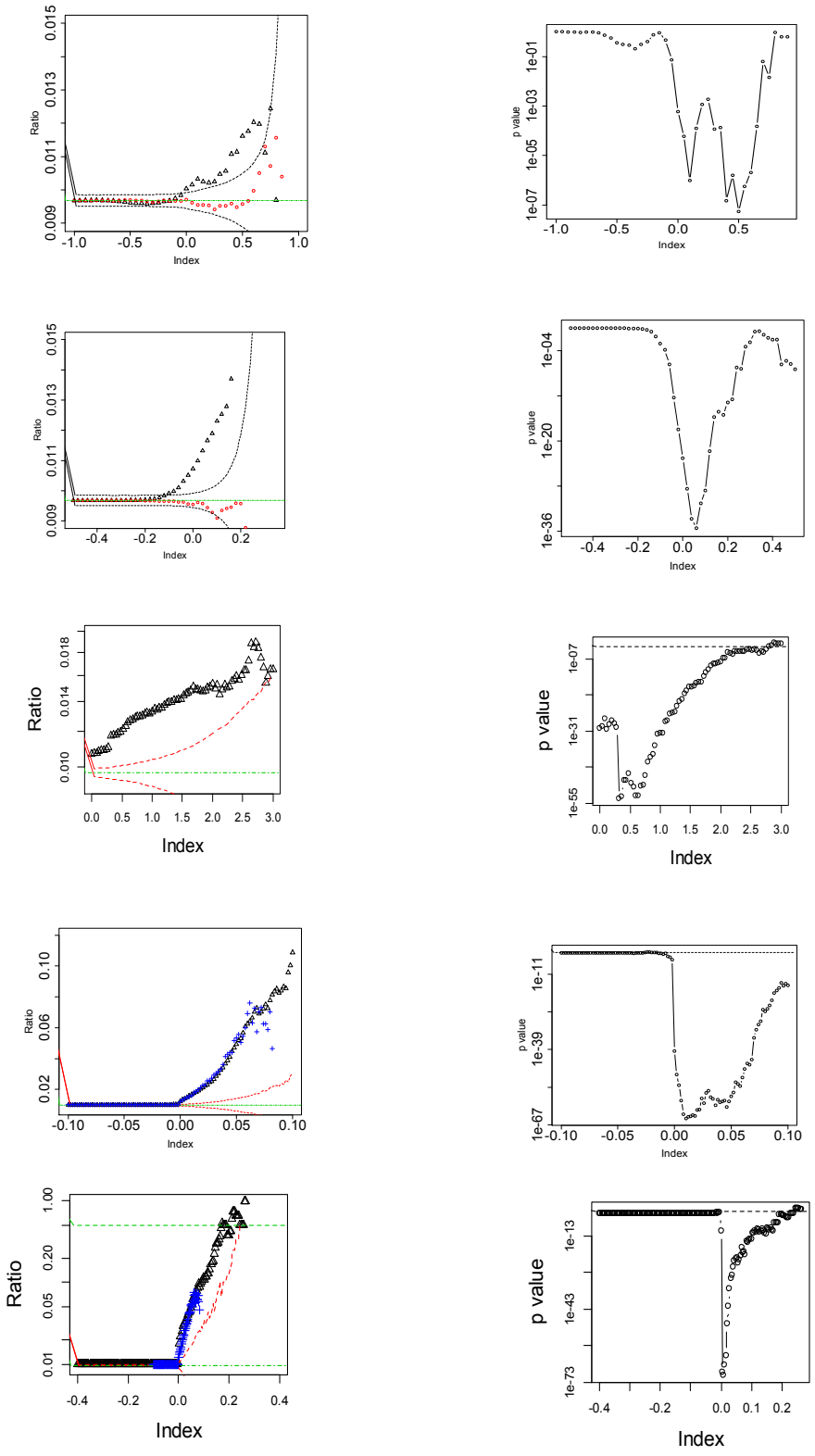

図 1: 相関指標の関係性抽出能力の比較（詳細は式 1 を参照.）．左：比率 $q(r)$（関係性の検出能力の指標）．点線は口コミ時系列をランダムシャッフルした時系列に置き換えて実験した場合（ランダムな場合）の１%線と９９%線．破線はランダムに選んだ場合の比率の期待値．右：二項検定の p 値 (ランダムとの相違の指標)．一段目：順位相関係数（赤丸は口コミ時系列をランダムシャッフルした場合）． 二段目：差分ボックス平均（株価終値），三段目：異常値相関，四段目：複合指標（平均差分ボックス相関：価格＋取引量）; 青十字は同様の指標を１４日にボックスサイズの時間平均（式４）に置き換えたもの． ５段目：複合指標:（平均差分ボックス相関：価格＋平均差分ボックス相関：取引量＋異常値）．一段目に示した順位相関ではほぼランダム選択に近く，二段目，三段目の単独の判定基準では，関係性性の検出力は一段目とあまり変わらないがランダムとの相違 (p 値) はより大きくなっていることがわかる．また，四段目，五段目の複合指標では一段目，二段目，三段目に示した結果に加えて検出力，ランダム選択との相違 (p 値) ともにが大きくなっていることがわかる．