D-013

株式掲示板を用いた投稿数のバースト発生時における株価動向・株式関連指標の予測

Predicting Stock Price and Related Indexes focusing on Bursts in Streams of Stock Bulletin Board

桜田 亮太† 青野 雅樹‡
Ryota Sakurada Masaki Aono

1. はじめに

近年，インターネットによる株式取引が普及した影響で株式市場において個人投資家は増加傾向にあり，中でも新興市場においては個人投資家が株価の相場形成に重要な役割を果たしている．個人投資家は各種銘柄の情報交換の場としてインターネットの株式掲示板を用いる事が多い．株式掲示板では情報交換が盛んに行われ，重要な事象が発生した際には掲示板の投稿数が跳ね上がり(バースト)，今後の動向について活発に議論されることが多い．

本研究では，新興市場銘柄における株式掲示板の投稿においてほぼ毎日いずれかの銘柄で観測されるバーストに着目し，投稿数のバーストを検知した際に投稿内容や投稿数などの情報を用いてバースト後の株価や売買代金(出来高)の動向，その他株式関連指標の動向の予測を行い，個人投資家の発言と株式市場動向との関連性を明らかにする．

2. 関連研究

関連研究として，掲示板など個人の投稿と株式の関係に着目した代表的な研究を述べる．

Antweiler(2004)[1]らは，ダウ・ジョーンズ指数に採用されている 45 社について掲示板の投稿数および内容を自然言語処理により分析し，掲示板は株式リターンを予測できないこと，投稿は出来高を予測できること，投稿はボラティリティを予測できることなどを指摘している．

Bollen(2010)[2]らや Vu(2012)[3]らは，Twitter の投稿を感情分析することで，株式指数や個別銘柄の株価を予測する研究を行っている．Bollen らは 3 日後の「ダウ・ジョーンズ工業株価平均」の変化の方向性を 86.7%の精度で，Vu らは Apple など IT 企業 4 社の翌日の始値を 4 社平均で 78.5%の精度で予測できると報告した．

3. 提案手法

本研究では，株式掲示板の投稿数が急激に上昇（バースト発生）した翌日の株価・売買代金を，株式掲示板の投稿内容や過去の株価・売買代金時系列データを用いることで予測する．提案手法の流れを図 1 に示す．

なお，株式掲示板のデータとしては Yahoo! JAPAN が提供する textream の株式カテゴリのデータを用いる．textream は株式掲示板としては日本最大級で銘柄ごとに掲示板が用意されていて議論がなされている．

提案手法の流れとして，まずそれぞれの銘柄について掲示板の投稿を用いて投稿数のバーストを検知し，バースト発生日の掲示板の投稿データ及び，バースト発生した翌日の株価・売買代金時系列データを取得する(3.1 節)．そして，説明変数を投稿数，目的変数を株価または売買代金として，それぞれの銘柄において回帰モデルを導出する．(3.2 節)

3.2 節において，銘柄によっては投稿数が少なく，モデル導出に必要な十分なデータが揃わない場合がある．それらの銘柄については，関連銘柄を株価時系列データから抽出し，その関連銘柄のバースト時のデータを用いて回帰モデルを導出する．関連銘柄の抽出方法は 3.3 節で述べる．

最後に，未知のデータで投稿数のバーストが発生した際に，3.2 節および 3.3 節で導出した回帰モデルを適用することで，翌日の株価・売買代金の予測を行う(3.4 節)．

3.1 投稿数のバースト検知

ある銘柄について投稿数が過去の投稿数と比較し急激に上昇することを”投稿数のバースト”と定義し，バースト検知モデルにより検出する．バースト検知は過去 10 日の移動平均投稿数との乖離率が閾値(150%)を超えた時とする．

なお，投稿数をデータとして用いると一人のユーザが一日の間に多量の投稿をした際に正常に予測できなくなる可能性があるため，投稿総数ではなく投稿したユーザの総数をデータとして用いる．

3.2 回帰モデル

バーストが発生した翌日の投稿データ及び株価・売買代金時系列データを用いて回帰モデルをそれぞれの銘柄について導出する．入力をバースト後の掲示板の投稿数，株価，または売買代金を予測する線形回帰モデルを最小二乗法により求める．

3.3 関連銘柄クラスタリング

掲示板の投稿数が十分でなく，正常に回帰モデルを導出できない場合が存在する．そのような場合には，投稿数の少ない銘柄と関連する銘柄を導出し，その関連銘柄を用いて投稿数の少ない銘柄の回帰モデルを作成する．

株式市場においては類似した銘柄は類似した株価の動きをすることが知られている．そこで，関連銘柄の抽出方法として，過去の株価の変動が類似した銘柄は株価や売買代金が同様の動きをすると仮定し，日々の株価の変動を用いてクラスタリングを行い抽出する．

関連銘柄の抽出時には潜在的ディリクレ配分法(Latent Dirichlet Allocation)を用い，トピック分類を行いそれぞれのトピックをクラスタとして定義する． LDA に入力するデータの例を表 1 に示す．なお表中の 4 桁の数字は各銘柄に与えられている銘柄コードを示す．

それぞれの銘柄について過去の株価推移から日々の騰落率を調査し，一日ごとに騰落率が閾値(+6%)を超えた銘柄のリストを作成し，それを LDA の入力として与える．

以上の操作で導出された関連銘柄を用いて，データ数が少ない銘柄において，同じクラスタの LDA における重みの大きい銘柄上位 20 銘柄の総データを用いて 3.1 節，3.2 節の操作を行い，回帰モデルを導出する．

†豊橋技術科学大学 大学院 情報・知能工学専攻
‡豊橋技術科学大学 情報・知能工学系

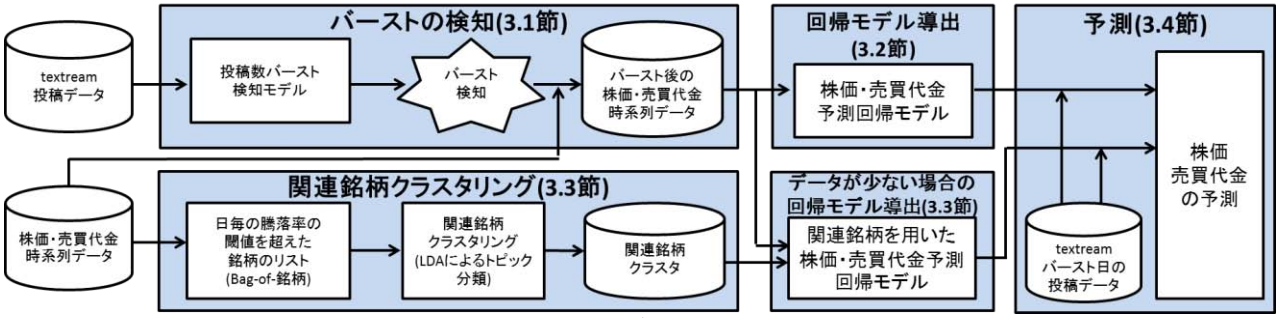

図 1　提案手法

3.4　今後の株価の予測

未知データについて掲示板投稿数を調査し，バーストが発生しているようであれば，3.2 節，3.3 節で導出した回帰モデルに適用し株価・売買代金を予測する．

表 1　LDA への入力データ例

5 月 1 日(文章 1)	9984　8515　7203　8306　8411　7011
5 月 2 日(文章 2)	8411　8515　8306　5401
5 月 3 日(文章 3)	4243　4575　2370

4．評価実験

提案手法有用性を評価するために実験を行った．本研究では個人投資家の多い新興市場銘柄を対象とし，東証マザーズ及び JASDAQ に上場している 1048 銘柄を用いて実験を行う．収集したデータは，2012 年 11 月 21 日から 2014 年 6 月 18 日までの 1048 銘柄の textream の投稿約 250 万件，またその期間のそれぞれの銘柄の株価時系列データである．訓練データとして 2012 年 11 月 21 日から 2014 年 4 月 30 日のデータ，テストデータとしてそれ以降のデータを用いる．

まず，関連銘柄クラスタリングの実験として，3.3 節の操作を行い関連銘柄のクラスタリングを行った．関連銘柄クラスタリングの出力結果の一部を表 2 に示す．

表 2　関連銘柄のクラスタリングの出力結果例

クラスタ 1	コロプラ，日本サード・パーティ，トランスジェニック，ジェイテック，そーせいグループ，エイティング…
クラスタ 2	プレシジョン・システム・サイエンス，ＤＮＡチップ研究所，ジーンテクノサイエンス，キャンバス，コスモ・バイオ…

出力結果例を見ると，クラスタ 1 のように異業種が関連銘柄としてクラスタリングされた例もあるが，クラスタ 2 のよう同業種の銘柄(クラスタ 2 はバイオ銘柄)が関連銘柄として抽出されるクラスタが多く見受けられた．

次に回帰モデルを用いて未知データの売買代金を予測した結果を示す．予測モデルの評価方法としては，平均相対誤差及び分類問題に帰着させた予測正解率を用いる．

平均相対誤差は予測値と実測値の当てはまりの良さを示す尺度であり，0 に近いほど良い回帰モデルと言える．

予測正解率は予測値の有用性を評価するための尺度で，予測値と実測値の誤差が±30%以内であれば正解とする分類問題に帰着させて評価する．

平均相対誤差 MRE，予測正解率 TPR は以下の式で表される．

$$\mathrm{MRE} = \frac{1}{n}\sum_{i=0}^{n}\frac{|y'_i - y_i|}{y'_i}$$

$$TPR = \frac{予測値と実測値の誤差が\pm 30\%以内の予測数}{総予測数} \times 100\ [\%]$$

ここで，n は実測値(予測値)の総数，y'_iは予測値，y_iは実測値，$\bar{y_i}$ は実測値の平均である．

提案モデルの評価を行うためにベースラインを選定し，回帰モデルとの評価尺度を比較し評価を行う．

ベースラインとして，バーストの有用性を調査するために，(1)バーストが発生していない場合を含めて投稿数と売買代金との回帰をとった場合，(2)バースト発生時のみ回帰をするが十分なデータがない銘柄に関連銘柄クラスタリングは適用せずすべての銘柄のデータを用いることで導出された回帰モデルを適用し予測する手法の 2 つを選定し，バースト及び関連銘柄クラスタリングを用いた場合とを比較し評価する．実験結果を以下の表 3 に示す．

表 3　ベースラインと提案手法の評価尺度の比較

	MRE	TPR
ベースライン(1)	2.1486	0.2168
ベースライン(2)	2.4652	0.2871
提案手法	2.2673	0.3366

結果を見ると，TPR で提案手法が他の手法を大幅に上回り，MRE においても 2 位という結果となった．また，投稿数のバーストが発生した際には 81%の精度で翌日の売買代金が上昇していることが判明した．なお売買代金は日々の変動幅が大きいため，MRE が大きい数値となってしまっている．この結果より，投稿数は売買代金の予測に有用な素性であること，バーストが発生時に特化させることで精度向上を実現でき，更に関連銘柄を用いることで十分なデータが揃わない銘柄における精度向上を実現できることが判明した．

6．おわりに

本研究では，株式掲示板で観測される投稿数バーストに着目し，投稿数のバーストを検知した際に投稿数を用いて株式関連指標の一種である売買代金の予測を行った．結果として，投稿数は売買代金の予測に有用な素性であり，バーストが発生時に特化させることで精度向上を実現でき，更に関連銘柄を用いることで十分なデータが揃わない銘柄における精度向上を実現できることがわかった．株価動向の予測については投稿数との相関がなく，予測が困難であった．今後の課題としては，売買代金だけでなく株価の予測，掲示板の投稿内容などを用いたテキストマイニングの予測への組み入れ，回帰モデルのオンラインでのパラメータの更新，関連銘柄の抽出の手法の改善などが挙げられる．

参考文献

[1] Antweiler W, Frank MZ, "Is all that talk just noise? the information content of internet stock message boards.", The Journal of Finance 59, 2004.

[2] Johan Bollen, Huina Mao, Xiao-Jun Zeng, "Twitter mood predicts the stock market. ", Journal of Computational Science 2 2011 1–8, 2011.

[3] Vu, T. T., Chang, S., Ha, Q. T., Collier, N. , "An experiment in integrating sentiment features for tech stock prediction in twitter. ", The COLING 2012 Organizing Committee, 2012.