高頻度金融データによる相互依存企業群の抽出

大坪　将之† JaroslavMelesko†† 江口　浩二†††

† 神戸大学工学部情報知能工学科　〒 657–8501 兵庫県神戸市灘区六甲台町 1-1
†† Vilnius Gediminas Technical University, Faculty of Fundamental Sciences, Department of Information Technologies　〒 LT–10223 Lithuania, Vilnius, Sauletikio al. 11
††† 神戸大学大学院システム情報学研究科情報科学専攻　〒 657–8501 兵庫県神戸市灘区六甲台町 1-1
E-mail: †otsubo@cs25.scitec.kobe-u.ac.jp, ††jaroslav.melesko@vgtu.lt, †††eguchi@port.kobe-u.ac.jp

あらまし　近年，フィンテックや仮想通貨などの普及が進みつつある中，金融時系列データの変動の解析や予測を行う研究に対する注目が高まっている．株価変動の予測についても長年研究が取り組まれており，様々な手法による株価予測が行われてきた．相互に依存する企業間では株価の挙動も互いに類似することが考えられるが，相互依存企業群を陽に記述したデータは入手が容易でない．そこで本稿では，株価歩み値データに基づいて企業間の依存性を抽出することを目的とする．本稿では 6 つの類似度算出方法によって株価変動において相互に依存する企業群の抽出を行い，各指標の精度の比較評価を行う．また，ニュース記事の発行時刻から n 時間後までの株価の変化を観測し，各指標による経過時刻による精度についても比較評価を行う．これにより，ニュース記事の発行時刻からの経過時間によって最適な指標について議論する．本稿では類似性指標として，(1) ピアソンの積率相関係数，(2) スピアマンの順位相関係数，(3) 正規化絶対残差，(4) 符号付絶対残差，(5) 補正積率相関係数, (6) 離散化相互情報量の 6 つの指標の比較を行った．これらの各指標の相互依存企業群の抽出精度を比較し，評価を行った結果，短期間においては補正積率相関係数が最も企業間の依存性を抽出することができ，長期間になるにつれて離散化相互情報量の抽出精度が高くなる傾向があることを示した．

キーワード　高頻度金融データ, 相互情報量，金融ネットワーク，株式市場予測

1. はじめに

近年，フィンテックや仮想通貨などの普及が進みつつある中，金融時系列データの変動の解析を行う研究に対する注目が高まっている．株価の変動の予測もその最たるものと言えるもので，長年研究が行われてきた．より良い精度で変動予測ができる手段を確立することによってビジネスへと展開することも可能になる．

こうした研究においては様々なアプローチから株価の予測が行われてきた．株式市場の金融指標を予測する目的で，株価の 1 時間の変動率を企業間の相互情報量 (Mutual Information: MI) を用いて時間ごとのネットワークを形成し，企業間の依存性を測る手法 [1] や，金融指標の上下を予測する目的で，企業が提出する報告書と，金融指標の関係性を解析し，学習することによって言語的特徴量と非言語特徴量を組み合わせたランダムフォレスト分類器による分類を行う手法 [2] などが提案されている．

本稿においては，企業間の株価変動の依存性を抽出するためのいくつかの指標を比較する．今回用いた類似度の尺度は以下のとおりである．

- ピアソンの積率相関係数
- スピアマンの順位相関係数
- 正規化絶対残差
- 符号付絶対残差
- 補正積率相関係数
- 離散化相互情報量

また，本稿ではニュース記事が発行された時刻からの一定期間の時間幅を変化させることによって記事の影響の持続時間，および各指標の解析可能時間について分析を行う．

本稿で示す手法を用いて企業間の依存性をより良く抽出することによって，実際に機関投資家や個人投資家達の助力となることが期待される．また，投資家達に使用されるだけでなく，今後，企業間の株価変動の依存性を考慮した株価予測の様々な研究の助力となることを期待する．実験では，ニュース記事が発行された時刻を起点とし，一定の時間幅において株価変動の動きが類似する企業間では依存性があると仮定し，いくつかの類似性指標を用いて相互依存企業群を抽出する．その評価には，ニュース記事で同時に言及された企業群を正解データとし，F 値（F-measure) [3] により評価する．ニュース記事データとして Thomson Reuters News（TRN）を用い，株価のデータとして東証株価歩み値データを用いる．

2. 関連研究

2.1 相互情報量による金融ネットワーク構築

文献 [1] では米国の代表的な平均株価指数である S&P500 の変動を予測するための手法のステップとして S&P500 に上場している企業の間の依存性を表す金融ネットワークを，個別企業の株価歩み値に基づく相互情報量を使用して構築している．相

互情報量は以下の式で定義される．

$$I(X_i;Y_j) = \sum_{x\in X_i, y\in Y_j, x\neq y} p(x,y)\log\frac{p(x,y)}{p(x)p(y)} \quad (1)$$

$$M = \begin{pmatrix} 0 & I(X_{s1};Y_{s2}) & \dots & I(X_{s1};Y_{sN}) \\ I(X_{s2};Y_{s1}) & 0 & \dots & I(X_{s2};Y_{sN}) \\ \vdots & \vdots & \ddots & \vdots \\ I(X_{sN};Y_{s1}) & I(X_{sN};Y_{s2}) & \dots & 0 \end{pmatrix} \quad (2)$$

ここで式 (1) はある企業間 (X_i, Y_i) の相互情報量を示している．相互情報量は非線形相関をより良くとらえることが可能である．各 X_i, Y_j は特定の時区間 (例えば 1 時間) の 1 分毎の株価，60 点からなる時系列データであり，$p(x)$ と $p(y)$ はそれぞれの相対頻度分布を示す．式 (2) はある時区分における対象企業の数を N とした時の全ての企業間における相互情報量からなる $N \times N$ 次の正方行列である．つまり，行列 M の 1 つの要素は X_i 社と Y_j 社の相互情報量を示しており，それぞれ i 行，j 列は特定の企業とその他の企業との相互情報量を示すベクトルを表す．ここでは株価そのものに基づいた相対頻度分布を用いているのに対して，本稿では株価の変動を直接捉えた類似性指標に着目する．

また，文献 [1] では，構築した金融ネットワークにより強度分布の変化をとらえることで S&P500 の市場指数の変動の予測を行っており，1 ステップ後の予測について有効であることが示されている．それに対して，本稿では市場指数ではなくネットワークを構成する個別銘柄について着目し，企業間の依存性の発見を行う．

2.2 LSTM を用いた金融時系列の深層学習

文献 [4] ではウェーブレット変換，積層オートエンコーダー (Stacked Autoencoder: SAE)，長短期メモリネットワーク (Long Short-Term Memory network: LSTM) を組み合わせた株価予測を行うフレームワークが提案されており，予測精度と収益性の両方の性能において類似モデルよりも優れていることが示されている．当文献では市場指数の変動について行われているが，LSTM にデータを入力する前処理としてウェーブレット変換などを行うことによってフレームワーク内で高周波成分のノイズ除去が行われることによって，精度が向上していることが示されている．つまり，本稿においては市場指数ではなく個別銘柄に着目するという点で文献とは異なるが，本稿で述べる相互依存企業群の抽出と LSTM を組み合わせることで，株価の予測性能をより高めることができると期待される．

2.3 テキストデータを用いた株価指標予測

文献 [2] では，これまで用いられていた株価の時系列データのみからでなく，S&P 500 に上場している企業のうち，各企業が提出する報告書のテキスト分析を行うことによって得られた言語的特徴量をさらにモデルに組み込んで予測を行い，該当企業の株価の変動予測を 3 値 (「UP」，「STAY」，「DOWN」) 分類で行っている．当文献では非言語的特徴量のみを用いた株価予測と比較して言語的特徴量を組み込んだモデルによって予測精度が向上することが示されている．本稿では，後述するようにテキストデータである Thomson Reuters News の記事は公開日時の取得と，該当記事に同時に記載のある企業を関連性のある企業群とみなす正解データの構築にのみ用いているが，将来的にはこれらの記事をトピックモデルや深層学習などによって解析することによって関連性のある企業群の株価予測を行うことが可能であり，株価予測の精度を向上できると期待される．

3. 提 案 手 法

本稿では，株価変動に基づいた相互依存企業群の抽出において，様々な類似度を計測する指標を用いたフレームワークを提案する．本稿の目的として，本稿で提案するフレームワークによって抽出された，株価変動について相互に依存する企業群の辞書の構築を行うことによって，将来的にある企業に対するニュース記事が発行された時，即座に関連企業群についても株の売買の判断を行うことが可能になると期待される．

提案するフレームワークのフロー図を図 1 に示す．

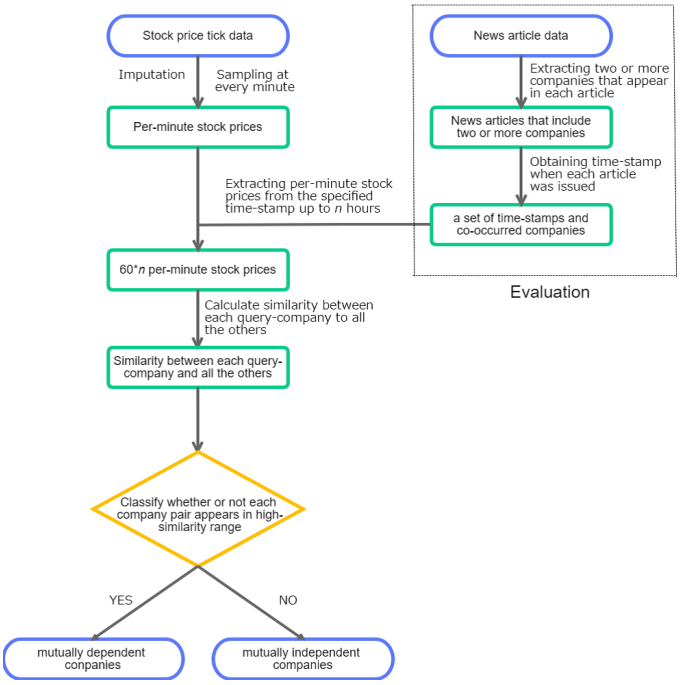

図 1 提案するフレームワークのフロー図

3.1 高頻度金融データによる株価情報の取得

着目する時刻から n 時間後までの高頻度金融データにおける 1 分間隔の株価の数値 $60 \times n$ 点を抽出する．後述するように高頻度金融データは取引が行われる毎に企業名や株価が記録されているデータであるが，本稿ではこれらを 1 分毎にサンプリングと補完を行うことによって 1 分間隔のデータとしている．本稿では高頻度金融データとして東証株価歩み値データを用いる．また，取得したい n 時間中に東京証券取引所 (東証) の休憩時間や，営業時間外の時間が含まれる場合，あるいは対象時間中で営業終了時刻を超える場合には，次の開始時刻まで繰り越してデータを抽出し，結合することで n 時間のデータとしてみなす．

高頻度金融データについては東証の営業開始時刻 09:00〜

15:00 において 1 分毎に該当時刻の直前の株価の値を抽出した．また，1 分間以上取引のない時刻，つまり欠損値に関してはホットデック法 (hot-deck imputation) により補完を行うことによってさらにその直前の値を現在の値とみなす．

3.2 相互依存企業群の抽出

着目する時刻から n 時間における，高頻度金融データから観測される株価変動の類似度を以下に述べるいくつかの指標を用いて算出する．各指標によって得られた類似度の頻度分布において一定基準以上の高類似度領域に存在する企業群と，正解データとして用いる実際に同時に記載されている企業群を比較し，各指標の抽出精度を測る．

また，以下に今回用いた類似性指標を示す．

3.2.1 ピアソンの積率相関係数

ピアソンの積率相関係数はあるデータ集合間の線形関係の強さを表す．データ $X = (x_1, \cdots, x_i, \cdots, x_N)$ とデータ $Y = (y_1, \cdots, y_i, \cdots, y_N)$ の相関係数 r_{xy} は以下の式で表される：

$$r_{xy} = \frac{\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{N}(y_i - \overline{y})^2}} \tag{3}$$

データ X,Y のデータ数はそれぞれ N 個であり，$\overline{x}$,$\overline{y}$ は，それぞれのデータ集合の平均を示す．ピアソンの積率相関係数は，$-1 \leqq r_{xy} \leqq 1$ で表され，相関が低いほど 0 に近づき，高ければ ± 1 に近い数値を示す．なお，$r_{xy} > 0$ は正の相関，$r_{xy} < 0$ は負の相関を示す．

3.2.2 スピアマンの順位相関係数

スピアマンの順位相関係数では，データ点数を N とする 2 つのデータ集合毎に数値によるランキング (1〜N 番) をつけ，その順位によって相関係数を算出する．スピアマンの順位相関係数 ρ_{xy} は以下の式で定義される：

$$\rho_{xy} = 1 - \frac{6\sum_{k=1}^{N} {d_k}^2}{N^3 - N} \tag{4}$$

ここで，d_i は 2 つのデータを，データ $X = (x_1, \cdots, x_i, \cdots, x_N)$ とデータ $Y = (y_1, \cdots, y_i, \cdots, y_N)$ として，着目する i 番目の点 x_i と y_i における各データの順位差を示す．また，同順位が生じる場合には，X，Y における同順位の個数をそれぞれ n_x，n_y，またそれらの順位を $t_i (i = 1, 2, \cdots, n_x)$,$t_j (j = 1, 2, \cdots, n_j)$ として，以下の式を用いる：

$$\rho_{xy} = \frac{T_x + T_y - \sum_{k=1}^{N} {d_k}^2}{2\sqrt{T_x T_y}} \tag{5}$$

$$T_x = \frac{N^3 - N - \sum_{i=1}^{n_x}(t_i^3 - t_i)}{12} \tag{6}$$

$$T_y = \frac{N^3 - N - \sum_{j=1}^{n_y}(t_j^3 - t_j)}{12} \tag{7}$$

ピアソンの積率相関係数と同様に，$-1 \leqq \rho_{xy} \leqq 1$ で表され，相関が低いほど 0 に近づき，高ければ ± 1 に近い数値を示す．

3.2.3 正規化絶対残差

株価のデータ点の距離による相関を計る指標として正規化絶対残差 (Normalized Absolute Difference: NAD) を導入する．NAD では 2 つのデータ集合をそれぞれ最小値を 0，最大値を 1 として正規化を行う．正規化して得られるデータ集合 $A' = (A'_1, \cdots, A'_i, \cdots, A'_N)$ はデータ集合 $A = (A_1, \cdots, A_i, \cdots, A_N)$ のうち，最大値をとるデータを A_{max}，最小値をとるデータを A_{min} とした場合に以下の式で表される．

$$A'_i = \frac{A_i - A_{min}}{A_{max} - A_{min}} \tag{8}$$

まず，正規化して得られたデータ集合における各データ間の距離の差の絶対値の総和を計算し，距離 d'_{xy} とする．d'_{xy} は，正規化を行った 2 つのデータ集合を $X = (x_1, \cdots, x_i, \cdots, x_N)$, $Y = (y_1, \cdots, y_i, \cdots, y_N)$ とした時に，以下の式で表される．

$$d'_{xy} = \sum_{i=1}^{N} |x_i - y_i| \tag{9}$$

ただし，NAD では一方のデータを反転させたものとの距離も測ることによって負の相関関係にある企業間についても抽出を可能にする．ここで，反転させたデータ $-A$ とは $-A = (-A_1, \cdots, -A_i, \cdots, -A_N)$ とする．また，$-A$ を同様に正規化を行ったものを Y' とし，式 (9) によって $(X, Y), (X, Y')$ についてそれぞれ d'_{xy} を算出し，より小さい方を NAD とする．

上式の d'_{xy} では数値が小さいほど相関が大きく，数値が大きければ相関は小さいことを示す．

3.2.4 符号付絶対残差

NAD では相関の正負を表現することは不可能である．そこで符号付絶対残差 (Signed Absolute Difference: SAD) を導入する．SAD では NAD において一方を反転させたものとの距離の差の総和の方が小さい場合，負の距離として扱うことで相関の正負を表現する．さらに，SAD d_{xy} はデータ数を N として式 (9) で得られる d'_{xy} を用いて以下の式で表される．

$$d_{xy} = 1 - \frac{d'_{xy}}{N} \tag{10}$$

つまり，一方を反転させたものとの距離の差の総和の方が小さい場合 SAD は $-d_{xy}$ とする．よって SAD の値は $[-1, 1]$ で定義され，相関が低いほど 0 に近づき，高ければ ± 1 に近い数値を示す．

3.2.5 補正積率相関係数

3.1 節で示したピアソンの積率相関係数を基本としつつ，さらに金融時系列の特徴を考慮した類似性指標として，補正積率相関係数 (Pearson correlation coefficient with LinerRegression: PLR) を提案する．データ X, データ Y のそれぞれの回帰直線の傾きを s_x，s_y とした時に，補正積率相関係数は 3.1 節で示したピアソンの積率相関係数 r_{xy} の大きさに s_x と s_y をかけあわせたものであり，以下の式で表される：

$$\lambda_{xy} = s_x \cdot s_y |r_{xy}| \tag{11}$$

これによって，回帰直線の傾きの絶対値が大きいとき，すなわ

表 1　離散化相互情報量の分割境界

discrete value	range
Super Down	$[-\infty, \mu-\sigma]$
Down	$(\mu-\sigma, -2(m-\mu))$
Stay	$[-2(m-\mu), 2(m-\mu)]$
Up	$(2(m-\mu), \mu+\sigma)$
Super Up	$[\mu+\sigma, \infty]$

ち株価の変化率が大きいときに大きな値をとり，そうでないときに小さな値をとる．さらに，データの回帰直線の傾きの符号が逆になっているデータ間での相関係数値は負の値となり，傾きの符号が同じ場合では相関係数値は正の値をとる．

3.2.6　離散化相互情報量

ここでは 2.1 節で示した相互情報量 (Mutual Imformation: MI) を用いる．ただし，2.1 節では株価そのものに基づいた相対頻度分布を用いているのに対して，本稿では株価の変動を直接捉えた類似性指標に着目する．本稿ではまず，各データ集合の各データ点の変化率を 5 段階 (「Super Down」，「Down」，「Stay」，「Up」，「Super Up」) に分類する．その際の分類の基準は, 着目する n 時間の企業の全データにおける株価変化率の平均 (μ)，中央値 (m), 標準偏差 (σ) を算出し，以下の表 1 の通りとする．

以上から分類された 5 段階の変化率によってヒストグラムを作成し，その相対頻度に基づく相対頻度分布とすることで式 1 を用いて相互情報量を算出する．相互情報量は相関が大きいほど数値は大きくなり，相関が小さいほど数値も小さくなる．

3.3　ニュース記事におけるタイムスタンプの取得

TRN のニュース記事において，特定の記事に 2 社以上が同時に記載のある記事に着目し，それらの記事のタイムスタンプを取得する．ここで得られたタイムスタンプの時刻を起点とした n 時間後までの株価の変動に関して 3.1 節と 3.2 節で述べた手法で分析し，評価する．

4. 実　　験

4.1　データセット

本稿では 2016 年 4 月〜2017 年 3 月における，株価歩み値データやティックデータと呼ばれる株の売買を記録した高頻度金融データを用いる．ここでは東京証券取引所 (東証) 上場企業について，取引が行われる毎にその時刻や株価を記録したデータを指す．これを 1 分毎にサンプリングすることによって n 時間の価格の変動を $60 \times n$ 個のデータによって観測する．東証営業時間 09:00〜15:00 における 1 分間隔で，該当時刻の直前の株価の数値をその時刻の数値とみなした．その際 1 分間以上取引のないデータ，つまり欠損値についてはホットデック法を用いて補間した．これは欠損値の直前の値を着目する時刻のデータとみなすものである．実験に用いる企業については，東証上場企業のうち，後述するロイター通信 (TRN) の 2016 年 4 月〜2017 年 3 月の記事において，1 度でも記載のある企業のみに着目する．

指標の評価に使用するテキストデータには，Thomson Reuters News(TRN) のニュース記事を用いる．TRN のニュース記事には発行日時 (タイムスタンプ)，ヘッドライン，記事本文，登場する企業とその銘柄コードなどが記載されている．ニュース記事は 2016 年 4 月〜2017 年 3 月の期間に発行されたものを実験データとして用いる．また，この内 2016 年 4 月〜2016 年 12 月までの記事を用いて指標毎にパラメータを決定する．さらに，決定したパラメータを用いて 2017 年 1 月〜2017 年 3 月の記事でテストを行う．TRN のニュース記事において，2 社以上の企業が同記事に登場する記事にのみ着目し，1 つの記事に同時に登場する企業においては関連があるものとみなして正解データとして用いる．また，各指標の相互依存企業群の抽出実験では，入力データとしてそれぞれ訓練データでは 2016 年 12 月，テストデータでは 2017 年 3 月を除いている．これは，2016 年 12 月を訓練データで入力した場合には 2017 年 1 月のデータを観測する可能性があり，テストデータと重複してしまう可能性があるためであり，同様にテストデータにおいては 2017 年 3 月を入力データとした場合に 2017 年 4 月のデータを観測する必要がある可能性があるためである．

4.2　実験における課題設定

TRN のニュース記事において 2 社以上の企業が同時に登場する記事の発行時刻より n 時間後までの，該当企業 1 社とその他全企業間における株価変動の類似度を算出しし，得られた分布においてある一定基準以上の高類似度領域に，同時に登場した残りの企業が存在しているかを観測する．また，高類似度領域の基準を変化させることによって各指標の抽出精度の比較を行った．この時の境界 B を得られた各分布の期待値を μ, 標準偏差を σ とした時に $B = \mu \pm \alpha \cdot \sigma$ とした．また，α は $[0.6, 2.0]$ の範囲で 0.6 から 0.1 の刻み幅とした．

本稿では指標による精度の比較だけでなく，ニュース記事が株価変動に与える影響の持続時間も考慮するため，これまで述べてきたようにニュース記事の発行時刻から株価変動の観測を行う n 時間を $n = 4^m (m \in \{0, 1, 2, 3\})$ と変化させることによってニュース記事の影響の持続時間も考慮した実験を行った．

4.3　評 価 方 法

評価尺度には F 値 (F-measure) [3] を用いる．データセットとして用いる TRN のニュース記事において，複数の企業が同時に記載されている記事に着目し，特定記事についての各指標の F 値を算出し，月ごとに平均を算出する．これによって得られた 2016 年 4 月〜2016 年 11 月までの 8ヶ月分の F 値の平均をさらに平均値を算出することによって各指標のスコアとし，比較する．各指標による抽出結果と，真の結果に基づいて表 2 のように分類する．

表 2　抽出結果と真の結果に基づいた分類

	true (actual)	false (actual)
positive (predicted)	true positives (tp)	false positives (fp)
negative (predicted)	false negatives (fn)	true negatives (tn)

F 値は，実験による抽出精度，再現率，それらの調和平均を

それぞれ

- 精度 (precision): 抽出したデータのうち，実際に真であるものの割合，つまり，一定以上の相関を持つとして抽出した企業のうち，実際に記事に同時に登場しているものの割合：

$$P = \frac{|tp|}{|tp| + |fp|} \tag{12}$$

- 再現率 (recall): 実際に真であるもののうち，抽出できたものの割合，つまり，実際に記事に同時に登場した企業のうち，一定以上の相関を持つとして抽出した企業の割合：

$$R = \frac{|tp|}{|tp| + |fn|} \tag{13}$$

- F 値 (F-measure)： 精度と再現率の調和平均:

$$F = \frac{2}{\frac{1}{P} + \frac{1}{R}} \tag{14}$$

さらに，精度に重みをおいた実験を行い，精度を重視した各指標の抽出精度についても比較する．この時の F 値は，$\beta^2 = \frac{1-\gamma}{\gamma} \in [0, \infty]$ として以下の通りとなる：

$$F_\beta = \frac{1}{\gamma \cdot \frac{1}{P} + (1-\gamma)\frac{1}{R}} = \frac{(1+\beta^2) \cdot P \times R}{\beta^2 \cdot P + R} \tag{15}$$

また，2016 年 4 月〜2016 年 11 月のニュース記事と高頻度金融データによって各指標で最も F 値の高かったパラメータ α を用いて，2017 年 1 月〜2017 年 2 月のデータをテストセットとしたテストを行う．

4.4 実験結果

パラメータ α を変化させた時の各指標の F 値の変化の一例 $(m = 0, m = 3)$ をそれぞれ図 2 および図 3 に示す．また，各 m の条件下で最も F 値の高い α と F 値を表 3 に示す．

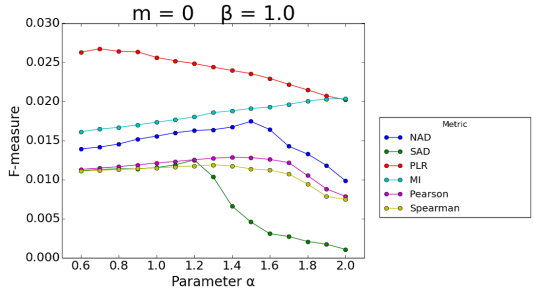

図 2　パラメータ α を変化させた時の F 値 $(m = 0, \beta = 1.0)$

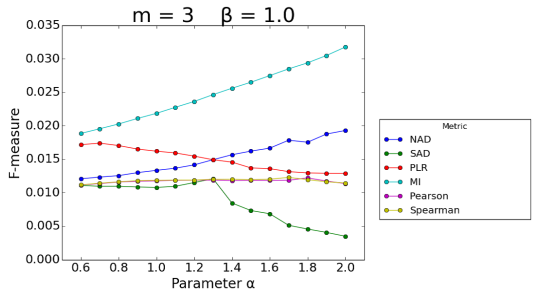

図 3　パラメータ α を変化させた時の F 値 $(m = 3, \beta = 1.0)$

表 3　各指標による F 値とパラメータ $\alpha(\beta = 1.0)$

Metric	$m=0$		$m=1$		$m=2$		$m=3$	
	α	F 値	α	F 値	α	F 値	α	F 値
NAD	1.5	0.0174	1.5	0.0186	2.0	0.0178	2.0	0.0193
SAD	1.2	0.0125	1.2	0.0132	1.2	0.0118	1.3	0.0121
PLR	0.7	0.0267	0.6	0.0243	0.6	0.0209	0.7	0.0174
MI	2.0	0.0203	2.0	0.0230	2.0	0.0284	2.0	0.0317
Pearson	1.4	0.0128	1.5	0.0144	1.7	0.0140	1.8	0.0122
Spearman	1.3	0.0119	1.3	0.0129	1.7	0.0135	1.7	0.0123

表 4　各指標による F 値とパラメータ $\alpha(\beta = 0.3)$

Metric	$m=0$		$m=1$		$m=2$		$m=3$	
	α	F 値	α	F 値	α	F 値	α	F 値
NAD	1.6	0.0145	2.0	0.0202	2.0	0.0159	2.0	0.0174
SAD	1.3	0.0079	1.2	0.0089	1.3	0.0073	1.3	0.0095
PLR	1.5	0.0200	1.0	0.0175	1.2	0.0145	0.8	0.0119
MI	2.0	0.0137	2.0	0.0146	2.0	0.0175	2.0	0.0194
Pearson	1.7	0.0086	1.7	0.0092	1.9	0.0096	1.8	0.0082
Spearman	1.7	0.0070	1.7	0.0080	1.7	0.0087	2.0	0.0080

表 5　各指標による F 値とパラメータ $\alpha(\beta = 0.5)$

Metric	$m=0$		$m=1$		$m=2$		$m=3$	
	α	F 値	α	F 値	α	F 値	α	F 値
NAD	1.6	0.0144	1.9	0.0181	2.0	0.0158	2.0	0.0171
SAD	1.2	0.0085	1.2	0.0095	1.2	0.0080	1.3	0.0095
PLR	0.9	0.0211	0.6	0.0188	0.6	0.0154	0.7	0.0130
MI	2.0	0.0150	2.0	0.0162	2.0	0.0196	2.0	0.0218
Pearson	1.7	0.0091	1.7	0.0101	1.8	0.0102	1.8	0.0089
Spearman	1.7	0.0077	1.7	0.0088	1.7	0.0096	2.0	0.0087

表 3 の結果から短期間の株価変動を観測する場合には PLR が最も相互依存企業群が抽出されていることがわかる．また，長期間にわたって観測をする場合では，期間が長くなるにつれて MI が最も相互依存企業群を抽出できることがわかった．これは，MI については株価の変化率を 5 段階に分けることによって類似性を測っていることによって，株価のデータ数が多くなるにつれて情報量も多くなり，より類似性を抽出できると推測される．さらに，PLR は長期間になるにつれて F 値は下がる傾向にある．これは，PLR は各データの回帰直線の傾きを考慮しているため，長期間になるにつれて傾きが 0 に近づくことが原因であると推測される．

また，上記の結果は精度 (precision) と再現率 (recall) に重みを与えていないが，本稿は実際に投資に役立てられることを目的としているため，精度が重視された F 値による比較も行った．本稿では式 (15) における β を，$\beta \in \{0.3, 0.5, 1.0\}$ と変化させることで異なる評価尺度として用いることで実験を行った．$\beta = 1.0$ の時は式 (15) は式 (14) と等価となる．よって結果は表 3 と一致する．$\beta = 0.3$，$\beta = 0.5$ の条件下で α を変化させた時の最も F 値が高くなる α と F 値を表 4 および表 5 に示す．また，各指標の F 値の変化の一例 $(m = 0, m = 3)$ を図 4〜7 に示す．また，各 β，m の条件下で，最も F 値の高い α と F 値を表 4 および表 5 に示す．

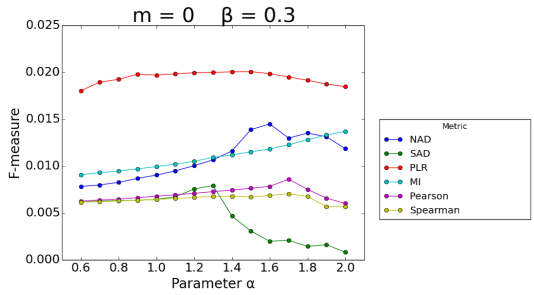

図 4　パラメータ α を変化させた時の F 値 ($m = 0$, $\beta = 0.3$)

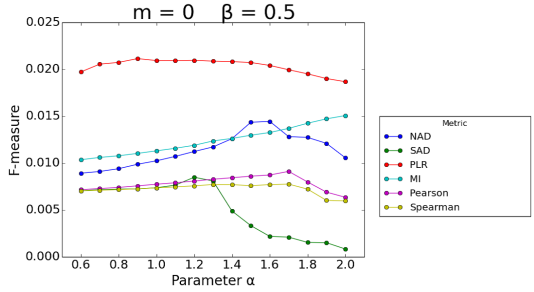

図 5　パラメータ α を変化させた時の F 値 ($m = 0$, $\beta = 0.5$)

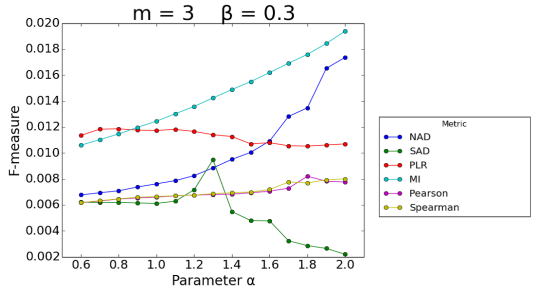

図 6　パラメータ α を変化させた時の F 値 ($m = 3$, $\beta = 0.3$)

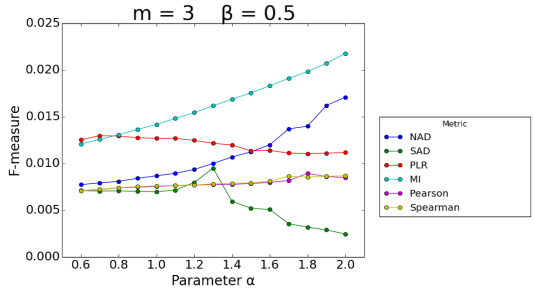

図 7　パラメータ α を変化させた時の F 値 ($m = 3$, $\beta = 0.5$)

次に以上の実験から得られた，指標毎に F 値を最大化する α を用いて，2017 年 1 月〜2017 年 2 月のデータにおいてテストを行った．この時の実験結果を以下に示す．

表 6　使用したパラメータ α による F 値のテスト結果 ($\beta = 1.0$)

Metric	$m=0$		$m=1$		$m=2$		$m=3$	
	α	F 値	α	F 値	α	F 値	α	F 値
NAD	1.5	0.0199	1.5	0.0208	2.0	0.0211	2.0	0.0203
SAD	1.2	0.0168	1.2	0.0147	1.2	0.0150	1.3	0.0095
PLR	0.7	0.0280	0.6	0.0265	0.6	0.0240	0.7	0.0220
MI	2.0	0.0241	2.0	0.0251	2.0	0.0327	2.0	0.0395
Pearson	1.4	0.0158	1.5	0.0169	1.7	0.0157	1.8	0.0110
Spearman	1.3	0.0142	1.3	0.0156	1.7	0.0155	1.7	0.0118

表 7　使用したパラメータ α による F 値のテスト結果 ($\beta = 0.3$)

Metric	$m=0$		$m=1$		$m=2$		$m=3$	
	α	F 値	α	F 値	α	F 値	α	F 値
NAD	1.6	0.0183	2.0	0.0233	2.0	0.0204	2.0	0.0174
SAD	1.3	0.0086	1.2	0.0092	1.3	0.0086	1.3	0.0060
PLR	1.5	0.0222	1.0	0.0193	1.2	0.0169	0.8	0.0158
MI	2.0	0.0174	2.0	0.0166	2.0	0.0210	2.0	0.0252
Pearson	1.7	0.0122	1.7	0.0108	1.9	0.0110	1.8	0.0074
Spearman	1.7	0.0085	1.7	0.0094	1.7	0.0103	2.0	0.0064

表 8　使用したパラメータ α による F 値のテスト結果 ($\beta = 0.5$)

Metric	$m=0$		$m=1$		$m=2$		$m=3$	
	α	F 値	α	F 値	α	F 値	α	F 値
NAD	1.6	0.0182	1.9	0.0201	2.0	0.0197	2.0	0.0177
SAD	1.2	0.0119	1.2	0.0102	1.2	0.0112	1.3	0.0066
PLR	0.9	0.0210	0.6	0.0177	0.6	0.0163	0.7	0.0166
MI	2.0	0.0188	2.0	0.0183	2.0	0.0233	2.0	0.0280
Pearson	1.7	0.0013	1.7	0.0119	1.8	0.0122	1.8	0.0081
Spearman	1.7	0.0094	1.7	0.0104	1.7	0.0113	2.0	0.0069

5. 追 加 実 験

追加実験として，上記では記事の発行時刻より n 時間後までの株価の変動のみを観測した結果であるが，2016 年 5 月〜2016 年 11 月の記事において，記事の発行時刻より $m = 0$，つまり 1 時間において $\alpha = 1.5$，$\beta = 0.3$ とした時の PLR について，1 時間前と 1 時間後のそれぞれの PLR の変化率による類似度の変化率に着目した実験を行った．この結果を図 8 に示す．ここで，各記事において発行時刻より 1 時間前までの PLR の数値を S_b，1 時間後までの PLR の数値を S_a とした時に，変化率 λ は以下の式で表される．

$$\lambda = \frac{|S_a - S_b|}{S_b} \tag{16}$$

PLR の変化率が λ 以上である記事のみを入力データとして上記実験と同様に F 値を算出した結果である．この結果から，$\lambda = 0.1$ 付近で F 値が高くなっているのは記事が発行された時点では依存した企業の株価の変動がすでに起こっている企業群が多く，また $\lambda = 1.0$ 付近で F 値が高くなっているのは記事が発行されたことによって株価の変動に依存性が発生した企業群も一定数ある，ということが推測される．

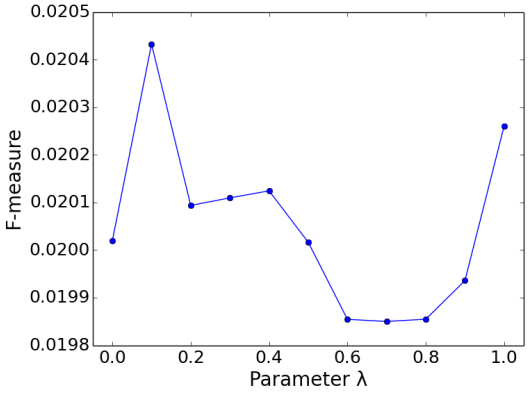

図 8　パラメータ λ を変化させた時の F 値 ($m = 0$, $\alpha = 1.5$ and $\beta = 0.3$)

6. おわりに

本研究では，企業の株価の変動に着目し，ニュース記事によって株価に受ける影響が依存する企業群の抽出を様々な類似性指標から分析を行った．今回用いた指標は以下の通りである．なお，このうち，後四者は本稿で新たに提案または修正を加えたものである．

- ピアソンの積率相関係数
- スピアマンの順位相関係数
- 正規化絶対残差 (NAD)
- 符号付絶対残差 (SAD)
- 補正積率相関係数 (PLR)
- 離散化相互情報量 (MI)

実験の結果から，ニュース記事の発行時刻からの短期間に着目すると，高頻度金融データによって相互依存企業群をより抽出することができる指標は PLR であり，長期間になるにつれて MI の抽出精度が向上することを示した．今後の課題としては，さらなる類似性指標を用いることができないか検討する必要がある．さらに，今回提案したフレームワークによって株価変動に着目した相互依存企業群の辞書を構築することによって，現実的に個人投資家達の役に立つツールとなることが期待できる．また，今回は α を 0.6〜2.0 の範囲で 0.1 の刻み幅の 15 個としたが，この α の探索範囲や，刻み幅などをさらに変化させることによってより最適なパラメータを発見し，各類似性指標の持つ性能の上限を探る必要がある．さらに，追加実験として行った λ の変化による結果を各指標のもとでパラメータを変化させた結果を探る必要がある．また，TRN の記事については公開日時の取得と，正解データの構築にしか用いていないが，2 章でも述べたようにこのテキストデータ自体をトピックモデルや深層学習などによって解析することによって，今回の研究結果を生かした相互依存企業群に着目した株価予測が可能であるという点に着目して，研究を進める必要がある．また，TRN のニュース記事に限らず，企業コードと公開日時が記載されているその他のニュース記事においても関連性があるかどうかを検討する必要がある．

謝　辞

本研究の一部は科学研究費補助金基盤研究 (B)（15H02703）の援助による．

文　献

[1] Minjun Kim and Hiroki Sayama. Predicting stock market movements using network science: an information theoretic approach. *Applied Network Science*, Vol. 2, p. 35, 2017.

[2] H Lee, M Surdeanu, B MacCartney, and D Jurafsky. On the importance of text analysis for stock price prediction. pp. 1170–1175, 01 2014.

[3] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA, 2008.

[4] Yulei Rao Wei Bao, Jun Yue. A deep learning framework for financial time series using stacked autoencoders and long-short term memory. *PLoS ONE*, Vol. 12, e0180944, , 07 2017.