株価とニュースの統合分析のためのヘテロトピックモデル

馬場　慧† 馬　強†

† 京都大学大学院情報学研究科　〒 606–8501 京都市左京区吉田本町
E-mail: †baba@db.soc.i.kyoto-u.ac.jp, ††qiang@i.kyoto-u.ac.jp

あらまし　株価の予測は，マーケティングや意思決定において重要である．ニュース記事を使用した株価の予測の研究は数多く存在する．本研究では，ニュース記事が株価に影響を与える可能性を考慮し，マルチトピックモデルを拡張して，数値データである株価とテキストデータであるニュース記事を統合的に分析するヘテロトピックモデルを提案する．提案モデルを用いれば，ニュースによる企業の株価の変動の予測と企業間の関係の統合分析が可能となる．

キーワード　グラフィカルモデル，トピックモデル，投資情報分析，時系列データ，関係マイニング

1. はじめに

株価を予測することは，マーケティングや意思決定において重要である．特に個人投資家に関しては，株価の予測を行う際に利用できる情報源は限られており，構造化されたデータである株価等の数値データや非構造化データであるニュース記事等のテキストデータは非常に重要なものとなっている．

ニュース記事から株価の予測をする研究は多く行われている．数値データに関しても，時系列解析等の手法を用いて，株価の変動を予測する研究は盛んに行われてきた．

株価は企業の価値を示している指標であり，企業の業績などを反映している．一方，ニュース記事は，投資判断の材料となるような企業の情報を示していることも多く，ニュース記事が株価に影響を与えている可能性は高い．我々の先行研究では，ニュース記事と株価データを統合分析することが企業間の明示的と暗黙的な関係の両方を明らかにすることが可能であることを示している [1]．先行研究では，我々は株価には企業の業績以外にも市場や業種の影響が含まれているとして合成モデルを提案し，ニュース記事の情報を用いて比較範囲の選定を行うことで，株価をそのまま比較するよりも正確な企業間の関係を導出することを確認できている．

本研究では，企業の競争力の分析や戦略決定，個人投資家の企業の成長性の分析や投資企業選定の支援を行うため，数値データである株価とテキストデータを用いて，ニュース記事を統合的に分析するヘテロトピックモデルを提案し，企業間の関係も含めた統合的な分析を行う．

ヘテロトピックモデルは，ニュース記事の単語はトピックが決定されて生成されると考えるトピックモデルである LDA(Latent Dirichlet Allocation) [2] を拡張した確率生成モデルである．出来事はニュース記事のトピック (内容) と株価の両方に影響を与える可能性が高いと仮定し，ベイズ推論に基づいてモデルを構築する．

本研究の主な貢献を以下にまとめる．

- ニュースによる企業の株価の変動の予測と企業間の関係の統合分析が可能となるモデルを提案している．
- 異なる特性を持ったテキストデータと数値データの両方を分析可能なヘテロトピックモデルを提案している．
- スイッチ変数を導入し，イベントの株価への影響の有無をモデリングできるようにしている．

本論文の構成は次の通りである．2 節ではテキスト情報が株価に与える影響や株価の予測に関する関連研究を示し，3 節では本研究で提案するモデルについて記す．4 節ではパラメータの調節について説明し，5 節では導出できると期待される結果の応用について述べる．そして，6 節は本研究のまとめである．

2. 関連研究

ニュース記事やソーシャルネットワーク等のテキスト情報が株価に影響を与えるという研究は多く存在する．Tetlock は Wall Street Journal の市場観測のコラム記事から悲観度を抽出し，ダウ工業平均株価と関係していることを明らかにした [3]．Bollen らは Twitter のテキスト情報であるツイートを解析し，世間のムードを測ることによってダウ工業平均株価の変動の予測する試みを行っている [4]．Garcia は 1905 年から 2005 年の間の New York Times の金融に関するニュースを分析し，ニュース記事の内容を用いた株価の予測が，景気が後退しているときに役立つことより，不況時において投資家の感情の影響がより顕著であることに言及している [5]．このようにテキスト情報は株価の変動を測るうえで重要な指標のひとつとなっている．本研究では，確率モデルを構築し，株価の変動を予測する．テキストデータの構造を分析するのではなく単語の生成確率を仮定したトピックモデルを提案する．

本研究では，トピックモデルとして Latent Dirichlet Allocation(LDA) [2] を用いることにより，ニュース記事の単語の潜在的トピックを決定する．トピックモデルを用いた研究は盛んに行われている．Rosen らは著者にトピックの確率分布が対応すると考え，LDA を拡張して，著者情報を含む文書の生成モデルを作成した [6]．Li らは Twitter のデータを用いて，ユーザに関するトピックとコミュニティに関するトピックを同時に発見するトピックモデルを提案し，ユーザの正確なトピックやコミュニティの焦点を明らかにしている．[7] しかしながら，これらの研究では分析するデータとしてテキストデータのみに焦点を当てている．本研究では，数値データである株価とテキス

トデータであるニュース記事のような性質の異なるデータを扱える点が異なる．

本研究で提案するヘテロトピックモデルでは，イベントによって影響を受ける企業群を推論でき，企業間の関係を分析できることにも期待している．企業は組織であり，組織と組織の関係ネットワークを分析する「組織間関係論」は長年研究の対象となっており [8]，現在も研究が行われている．我々の先行研究では，ニュース記事と株価データを統合的に分析することによって，企業間の明示的，暗黙的な関係の両方を明らかにすることが可能であることを示している [1]．金らは Web 上に存在している情報から企業間の関係を明らかにし，企業ネットワークを抽出する手法を提案している [9]．金らの研究は企業間の関係性を導き出す情報として，Web 上のテキスト情報のみを対象としており，抽出する関係性の対象も提携関係と訴訟関係のみに絞っているが，本研究では関係性の対象を取らず，企業の業績に焦点を当てて関係性があるかどうかを判断する．

Woo らは協調性，適応性，雰囲気といった観点から企業間関係の質を評価し，関係とサービスの質の関係を明らかにする手法を提案している [10]．また，Rauyruen らは企業間関係の質を決定する要因としてサービスの品質や売り手へのコミットメント，信頼性，満足度をあげており，関係の質と購買の意図，繰り返し買うかどうかの忠実性に及ぼす影響の関係を調べた [11]．Woo らや Rauyruen らは企業間の関係の質をサービスの向上や顧客分析に利用するものと位置づけており，意思決定には利用しない．

3. 提案モデル

本節では数値データである株価とテキストデータであるニュース記事を統合的に分析するヘテロトピックモデルについて述べる．ニュース記事が企業の株価に影響を与える可能性を考慮するため，観測データとして株価データも用いる，トピックモデルである LDA を拡張したヘテロトピックモデルを提案する．

3.1 トピックモデル

トピックモデルとは，確率モデルの一種であり，ある文書が複数のトピックの混合として成り立っていると仮定し，単語の出現確率を推定するモデルである．文書はトピックの混合分布，トピックは単語の混合分布として表現される．トピックモデルとして頻繁に用いられる LDA では，観測データとしてドキュメント内の単語，潜在変数として単語の持つ潜在トピックを導入している．

3.2 ヘテロトピックモデル

LDA をはじめとしたトピックモデルでは，観測データとして用いられるのはテキストデータのみであり，テキスト以外の観点を含めた多角的な分析を行えないという欠点がある．先にも述べた通り，企業のニュース等は株価に影響を与える可能性が高く，テキストデータのみの分析よりも，株価のデータを含めた分析の方がニュースが企業に与える影響を考慮し，企業間の関係等まで含めた統合的な分析を行える．図 1 で示しているのが提案するヘテロトピックモデルのグラフィカルモデルである．図中の変数を表 3.2 にまとめる．

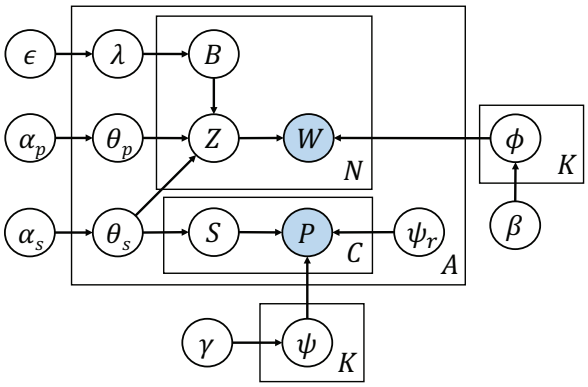

図 1　ヘテロトピックモデル

表 1　ヘテロトピックモデルの変数

変数	意味
A	ニュース記事の数
K	トピック数
N	ニュース記事 a の単語数
C	上場企業の数
B	ディリクレ分布を切り替えるスイッチ変数
S	トピック k が企業のグループにどのくらい影響を与えているかを示す潜在変数
Z	ニュース記事 a の n 番目の単語のトピック
W	ニュース記事 a の n 番目の単語
P	企業 c の株価の変動
α_p	企業の株価に影響を与えない単語に関するトピック分布であるディリクレ分布の事前ハイパーパラメータ
α_s	企業の株価に影響を与える単語に関するトピック分布であるディリクレ分布の事前ハイパーパラメータ
β	単語分布であるディリクレ分布の事前ハイパーパラメータ
γ	株価のニュースの影響による企業の株価の変動を示すコーシー分布の事前ハイパーパラメータ
ϵ	スイッチ変数の事前分布であるベータ分布の事前ハイパーパラメータ
θ_p	企業の株価に影響を与えない単語に関するトピック分布
θ_s	企業の株価に影響を与える単語に関するトピック分布
ϕ	トピック k の単語分布
ψ	企業 c の株価のニュースの影響による変動を示す分布
ψ_r	ニュース記事ごとに変動する価格変動バイアス
λ	スイッチ変数の分布

ヘテロトピックモデルの特徴を以下に列挙する．

- 観測データとして数値データである企業の株価の変動 P を導入する．
- トピック k が企業のグループにどのくらい影響を与えているかを示す，潜在変数 S を導入する．
- スイッチ変数 B を導入し，トピックの株価への影響の有無をモデリングする．ニュース記事内の単語のうち，企業の株価に影響を与えるものと与えないもので考慮する確率分布を切り替えることができる．

表 2　ヘテロトピックモデルのアルゴリズム

Prior
For each topic $k = 1, \cdots, K$:
　$\phi_k \sim Dirichlet(\beta)$
　$\psi_k \sim Cauchy(\gamma_r)$
End
For each article $a = 1, \cdots, A$:
　$\theta_{p_a} \sim Dirichlet(\alpha_p)$
　$\theta_{s_a} \sim Dirichlet(\alpha_s)$
　$\lambda_a \sim Beta(\epsilon)$
End

Likelihood
For each article $a = 1, \cdots, A$:
　$B \sim Bernoulli(\lambda)$
　For each word $n = 1, \cdots, N_a$:
　　If $B == 1$:
　　　$Z_{an} \sim Categorical(\theta_{s_a})$
　　　$W_{an} \sim Categorical(\phi_{Z_{an}})$
　　Else:
　　　$Z_{an} \sim Categorical(\theta_{p_a})$
　　　$W_{an} \sim Categorical(\phi_{Z_{an}})$
　End
　For each company $c = 1, \cdots, C$:
　　If $B == 1$:
　　　$S_{ac} \sim Categorical(\theta_{s_a})$
　　　$P_{ac} - \psi_r \sim Categorical(\psi_{S_{ac}})$
　End
End

- 一つのパラメータにトピック分布と企業のグループの分布の二つの意味を持たせることにより，特性の異なるテキストデータと数値データの二つの観測データを用いることを可能とする．
- ニュース記事ごとに価格変動バイアスを生成し，ニュース記事以外の影響による株価変動を考慮する．

表 3.2 にヘテロトピックモデルのニュース記事及び企業の株価の変動の生成アルゴリズムを示す．

各トピックの単語分布の事前分布 ϕ_k として，ディリクレ分布を仮定する．そして，企業の株価の変動の分布 ψ の分散は半コーシー分布から導出されるとしている．コーシー分布を仮定する理由としては，単なる正規分布を仮定した場合，外れ値に対応できないといった欠点があるためである．株価の変動では，企業の大きなニュースが発生したとき，大きな変動をすると予想されるので，外れ値に対応する必要がある．よって，裾が重くなっているコーシー分布を仮定する．スイッチ変数の分布の事前分布 λ として，ベータ分布を仮定する．スイッチ変数の分布としては，ベルヌーイ分布を仮定している．

$B = 0$ のとき，つまり，ニュース記事の単語 W_{an} が企業の株価に影響を与えない時，Z_{an} はトピック分布のカテゴリカル分布から導出される．このとき，W_{an} は単語分布のカテゴリカル分布から導出される．つまり，LDA と全く同じ処理を行う．

$B = 1$ のとき，つまり，ニュース記事の単語 W_{an} が企業の株価に影響を与える時，Z_{an} はトピック分布のカテゴリカル分布から導出される．このとき，同時に S_{ac} がグループ分布のカテゴリカル分布から導出される．S_{ac} はトピック k が企業 c にどのくらい影響を与えているかを示す潜在変数であるため，θ_s がトピックの分布であるとすると導出できない．

ここで，θ_s が二つの意味を持つパラメータであるとする．一つは先にも述べている通り，トピック分布である．もう一つの意味としては，企業のグループの分布という意味である．これは，$\theta_s = (G_1, G_2, \cdots, G_K)$ の K 次元ベクトル (K はトピック数) であるとした時，それぞれの要素は企業のグループを示している．企業のグループとは，そのトピックによって株価が変動する企業群である．企業のグループがカテゴリカル分布によって抽出されることによって，対応する企業の $\psi_{S_{ac}}$ の値を決定する．これは，Z_{an} が決定されると $\phi_{Z_{an}}$ が決定される LDA と同じような処理である．このように θ_s に 2 つの意味を持たせることにより，特性の異なるテキストデータと数値データを用いたモデルを作成することが可能となる．決定された $\psi_{S_{ac}}$ とニュース記事ごとに生成する価格変動バイアスの和によって企業の株価の変動 P_{ac} は生成される．

単語分布，単語のトピック分布，スイッチ変数の分布，株価の変動生成分布を統合的に考慮した分布は以下の式で表される．

$$\begin{aligned} &p(\mathbf{W}, \mathbf{P}, \mathbf{Z}, \mathbf{S}, \mathbf{B} | \alpha_p, \alpha_s, \beta, \gamma, \epsilon) \\ &= p(\mathbf{Z}|\mathbf{B}, \alpha_p, \alpha_s) p(\mathbf{W}|\mathbf{Z}, \beta) p(\mathbf{S}|\mathbf{B}, \alpha_s) p(\mathbf{P}|\mathbf{S}, \gamma) p(\mathbf{B}|\epsilon) \end{aligned} \tag{1}$$

4. パラメータの学習

本研究では実際のニュース記事を対象に，パラメータの学習を行う．本節では，使用するデータセット，パラメータを学習させる方法について述べる．

4.1 データセット

株価データは東京証券取引所 (注1) の歩み値データを使用する．歩み値には日中の約定値段の推移が記録されている．日足データではなく，歩み値を用いることによって，日毎の始値，終値だけでなく，いかなる瞬間の企業の株価の値でも取り出すことができる．ニュース記事が報道されたとき，即時に株価に反映されると考えられるため，日足データではなく，歩み値を用いる．対象とする銘柄としては，2015 年 1 月から 2016 年 10 月の間に一度でも取引が行われた銘柄とする．

ニュース記事のデータは，日経 QUICK(注2) のニュース記事を使用する．日経 QUICK は金融情報サービス会社 QUICK が提供する株式投資金融情報サイトであり，市場の動向を即座に反映している．ニュース記事が企業の株価に影響を与えているかどうかを調査するにあたって，ニュース記事がリアルタイムで掲載されるかどうかは重要な要素である．ニュース記事の構成要素として，ニュースコード，公開日時，タイトル，本文があるが，その中でも，株価の変動を導出する指標となる公開日時と単語の集合である本文を利用する．また，タイトルのみで

(注1)：http://db-ec.jpx.co.jp/
(注2)：http://www.quick.co.jp/page/top.html

内容を表しており，タイトルが存在しないフラッシュニュースに関しては，本研究では対象外とした．

4.2 パラメータ推定

本研究では，パラメータである θ_p，θ_s，ϕ，ψ，λ を推定する方法としてギブスサンプリングを用いた．ギブスサンプリングは LDA のパラメータ推定でも用いられる有名な手法である．ギブスサンプリングを用いてトピックの確率分布，スイッチ変数の確率分布，グループの確率分布を更新する．ギブスサンプリングの更新式を以下の式で定義する．

$$
\begin{aligned}
&p(b_i = 1|\mathbf{b_{-i}}, w, z, s, p) \\
&\propto \frac{p(b_i = 1, \mathbf{b_{-i}}, w, z, s, p)}{p(\mathbf{b_{-i}}, w, z, s, p)} \\
&\propto p(b_i = 1|z_i) \\
&\propto p(z_i|b_i = 1) \cdot p(b_i = 1) \\
&\propto \frac{n_{z_i,b_i=1} + \alpha_s}{\sum_{z_i} n_{z_i,b_i=1} + T\alpha_s} \cdot \frac{n_{b_i=1} + \epsilon}{\sum_{b_i} n_{b_i=1} + 2\epsilon}
\end{aligned} \tag{2}
$$

$$
\begin{aligned}
&p(b_i = 0|\mathbf{b_{-i}}, w, z, s, p) \\
&\propto \frac{p(b_i = 0, \mathbf{b_{-i}}, w, z, s, p)}{p(\mathbf{b_{-i}}, w, z, s, p)} \\
&\propto p(b_i = 0|z_i) \\
&\propto p(z_i|b_i = 0) \cdot p(b_i = 0) \\
&\propto \frac{n_{z_i,b_i=0} + \alpha_p}{\sum_{z_i} n_{z_i,b_i=0} + T\alpha_p} \cdot \frac{n_{b_i=0} + \epsilon}{\sum_{b_i} n_{b_i=0} + 2\epsilon}
\end{aligned} \tag{3}
$$

$$
\begin{aligned}
&p(z_i|\mathbf{z_{-i}}, w, b_i = 1, s, p) \\
&\propto \frac{p(\mathbf{z_i}, w, b_i = 1, s, p)}{p(\mathbf{z_{-i}}, w, b_i = 1, s, p)} \\
&\propto p(z_i, b_i = 1, w_i) \\
&\propto p(w_i|z_i) \cdot p(z_i|b_i = 1) \\
&\propto \frac{n_{w_i,z_i} + \beta}{\sum_V n_{w_i,z_i} + V\beta} \cdot \frac{n_{z_i,b_i=1} + \alpha_s}{\sum_{z_i} n_{z_i,b_i=1} + T\alpha_s}
\end{aligned} \tag{4}
$$

$$
\begin{aligned}
&p(z_i|\mathbf{z_{-i}}, w, b_i = 0, s, p) \\
&\propto \frac{p(\mathbf{z_i}, w, b_i = 0, s, p)}{p(\mathbf{z_{-i}}, w, b_i = 0, s, p)} \\
&\propto p(z_i, b_i = 0, w_i) \\
&\propto p(w_i|z_i) \cdot p(z_i|b_i = 0) \\
&\propto \frac{n_{w_i,z_i} + \beta}{\sum_V n_{w_i,z_i} + V\beta} \cdot \frac{n_{z_i,b_i=0} + \alpha_p}{\sum_{z_i} n_{z_i,b_i=0} + T\alpha_p}
\end{aligned} \tag{5}
$$

$$
\begin{aligned}
&p(s_i|\mathbf{s_{-1}}, w, b_i = 1, z, p) \\
&\propto \frac{p(\mathbf{s_i}, w, b_i = 1, z, p)}{p(\mathbf{s_{-i}}, w, b_i = 1, z, p)} \\
&\propto p(s_i, b_i = 1, p_i) \\
&\propto p(p_i|s_i) \cdot p(s_i|b_i = 1) \\
&\propto \frac{n_{p_i,s_i} + \gamma}{\sum_C n_{p_i,s_i} + C\gamma} \cdot \frac{n_{s_i,b_i=1} + \alpha_s}{\sum_{s_i} n_{s_i,b_i=1} + T\alpha_s}
\end{aligned} \tag{6}
$$

ただし，T はトピック数，V は全単語数，C は企業数，$n_{z_i,b_i=x}$ はスイッチ変数が x の値をとるときに文書がトピック z_i に割り当てられる回数，$n_{b_i=x}$ はスイッチ変数が x となるときの回数，n_{w_i,z_i} は単語 w_i がトピック z_i に割り当てられた回数，n_{p_i,s_i} は企業の株価の変動 p_i がグループ s_i に割り当てられた回数である．

本研究では，ある経済事象が発生したときにニュース記事の生成過程，それによって発生する株価の変動過程を明らかにすることを目的としている．よって，まず，ニュース記事において単語をトピックに割り当て，トピック分布のパラメータを更新する．次に，更新されたパラメータを用いて，ニュース記事が株価に与える影響を考慮し，株価の変動が類似しているグループの分布 (トピック分布) のパラメータを更新する．これを繰り返すことにより，ニュース記事が企業の株価に与える影響を考慮したヘテロトピックモデルのパラメータを学習する．

4.3 モデル評価

モデルの評価は株価の予測ができているかどうかを評価することによって行う．データの半分を教師データとし，パラメータの学習を行い，残りのデータでモデルの評価を行う．ニュース記事を入力とした時，学習済みのパラメータを用いて，株価の推定が行えているかを確認する．

5. 実　　験

本節では，実際の株価データとニュース記事データを用いてヘテロトピックモデルのパラメータを学習させた結果と考察を示す．

5.1 データの前処理

- 株価データ

本研究では，ニュース記事が報道されたときに，企業の株価の変動を生成するモデルを提案している．よって，このとき株価の値の大小ではなく，株価の変動率を導出しておく必要がある．変動率を用いて株価の変動を考えると，企業によって値の大きさが違う株価をすべて同じスケールのデータに変換できる．ここで，ニュース記事が株価に影響を及ぼす期間を 24 時間 (注3) とし，株価の変動率 δ を以下の式で求めた．ここで，ニュース記事が公開されたときの企業 c の株価を p_t，24 時間後の株価を p_{t+24} としている．

$$
\delta = \frac{p_{t+24} - p_t}{p_t} \tag{7}
$$

- ニュース記事データ

ニュース記事のデータとして，本文のデータを利用するが，ヘテロトピックモデルの入力として用いるために，単語に分割する必要がある．単語に分割するために，日本語形態素解析のソフトである Mecab [12] を利用する．Mecab に文章を入力すると，単語に分割され，品詞等が出力される．辞書としては，最新の固有名詞等も登録されている mecab-ipadic-neologd (注4) を用

(注3)：最適な期間を求める方法については，今後の研究で検討する予定である．
(注4)：https://github.com/neologd/mecab-ipadic-neologd/

表 3　各変数，各分布のハイパーパラメータ

変数・ハイパーパラメータ	値
A	10
K	36
V	371
C	3955
α_p	要素が全て 1 の K 次元ベクトル
α_s	要素が全て 1 の K 次元ベクトル
β	要素が全て 0.5 の V 次元ベクトル
γ	2.5
ϵ	(5, 3)

いた．本研究においては，株価に影響を与える可能性のある単語のみを考慮すべきである．よって，助詞や助動詞等のニュースの内容を表さない単語は無視し，株価に影響を与える可能性のある数字以外の名詞，形容詞のみをヘテロトピックモデルの入力とする．

5.2 考　　察

コーパスとして，日経 QUICK で公開されたニュース記事のうち，2015 年の 1 月の市場が公開される午前 9 以降の 10 記事を利用する．トピック数は東証が設定している 33 業種と同じ数に設定した．各変数の値，各分布のハイパーパラメータの値を表に示す(注5)．

この条件下で実験を行なったが，ニュース記事が 10 記事と少数であるにも関わらず，実行時間が非常に長いものとなっており，実用性があるとは言えない結果となった．

今回は，トピック数を 10 に減らすことで，学習を行なった結果について考察する．トピック数を削減したことにより，実行時間の問題は改善された．しかし，トピック数を減らすことは，企業のグループ数を減らすことと同義である．トピック数を増やすことで，よりミクロなニュース記事のクラスタリングを行うことができ，企業のグループの粒度を小さくすることに期待している．

ϕ の値に関しては，トピックが異なっても，同じような値に収束していることもある．原因としては，データが少なくてトピックの分離が適切にできていないことが考えられるため，ニュース記事の件数を増やし，トピックから単語が生成される確率を再度計算し直すことが課題である．

ψ の値はトピックごとに企業の株価の変動率を導出する分散を示している．値には分散が限りなく 0 に近い時が存在するが，コーシー分布の中央値を 0 としているため，この ψ の値をとるような株価の変動は限りなく 0 に近いものと考えられる．マーケットに関するニュース記事が報道されたとき，多くの企業が反応することも考えられるので，データ量を増やした確認が求められる．

最後に λ の値に関してであるが，経済に関するニュースに絞って学習させたにもかかわらず，最大でも 0.6 程度となっており，主に θ_s を経由して単語が生成されるとは言えない．この場合も，データ量を増加し，経済に関するニュースが実際に株価に影響を与えているのかどうかを調査する必要がある．

株価変動の推定周辺で離散確率分布と連続確率分布が混在しており，パラメータ学習が困難である可能性がある．大規模なデータセットで実験が行えるようにモデルの実装を見直す必要がある．

6. 応　　用

6.1 株価変動の推定

本研究では，企業の株価の変動を推定する確率生成モデルを提案している．提案モデルを用いると，ニュース発生時においてニュース記事に複数の企業が出てきた場合，その中から投資先企業の選定を行う支援ができる．また，ニュース記事に直接言及されていない企業でも，ニュース記事に言及されている企業との関係性があると判明していると，投資先の候補に入るといったことも考えられる．アプリ等を作成することができれば，企業や個人投資家などが投資先選定の際の意思決定の支援を行う効果が期待できる．

6.2 企業間の関係分析

本研究では，提案モデルに潜在変数を導入し，企業間の関係も導出できることを期待している．分析結果の提示は，企業の競争力の分析や意思決定の支援を行うことができる．投資の観点から見ても，ニュースで言及された企業の関連企業への投資を行うことで，影響が伝播するタイムラグを利用した投資が行えることに期待できる．

7. おわりに

本研究では，数値データである株価とテキストデータであるニュース記事を併用した株価の変動の予測と企業間の関係の統合分析を行うための確率生成モデルであるヘテロトピックモデルの提案を行っている．また，実際のデータを用いて，パラメータの学習を行なった．提案モデルを用いると，ニュース発生時における投資先企業の選定，企業や個人投資家などによる企業間ネットワーク分析の際の支援を行う効果が期待できる．

今後，行なっていく予定を以下に示す．

- 大規模なデータセットで実験が行えるように，モデル，実装の見直しを行う．
- 学習させるニュース記事のデータを増やし，より多くの場合を考慮できているパラメータを推定する．
- 推定されたパラメータを用いて実際のニュース記事が生成された時の株価の変動を推定し，モデルの正当性の評価を行う．
- 初期値によるパラメータ推定の差異を調査し，モデルにより適した初期値を求める．

モデルの正当性が確認されれば，結果を具体的に利用できる方法を考え，サービスとして投資家等が利用できるものを制作する予定である．

(注5)：最適な初期値を求める方法については，今後の研究で検討する予定である．

謝　　辞

本研究の一部は，科研費（課題番号 25700033) による．

文　　献

[1] Baba, S. and Ma, Q.: Analyzing Relationships of Listed Companies with Stock Prices and News Articles, *DEXA 2016*, LNCS 9827, pp. 27–34 (2016).

[2] Blei, D. M., Ng, A. Y. and Jordan, M. I.: Latent dirichlet allocation, *Journal of machine Learning research*, Vol. 3, No. Jan, pp. 993–1022 (2003).

[3] Tetlock, P. C.: Giving Content to Investor Sentiment: The Role of Media in the Stock Market, *The Journal of Finance*, Vol. 62, pp. 1139–1168 (2007).

[4] Bollen, J., Mao, H. and Zeng, X.: Twitter mood predicts the stock market, *Journal of Computational Science*, Vol. 2, pp. 1–8 (2011).

[5] Garcia, D.: Sentiment during recessions, *The Journal of Finance*, Vol. 68, No. 3, pp. 1267–1300 (2013).

[6] Rosen-Zvi, M., Griffiths, T., Steyvers, M. and Smyth, P.: The author-topic model for authors and documents, *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, AUAI Press, pp. 487–494 (2004).

[7] Li, L., Peng, W., Kataria, S., Sun, T. and Li, T.: Recommending Users and Communities in Social Media, *ACM Trans. Knowl. Discov. Data*, Vol. 10, No. 2, pp. 17:1–17:27 (2015).

[8] 山倉健嗣: 組織間関係と組織間関係論, 横浜経営研究, Vol. 16, No. 2, pp. 166–178 (1995).

[9] 金英子, 松尾豊, 石塚満: Web 上の情報を用いた企業間関係の抽出, 人工知能学会論文誌, Vol. 22, pp. 48–57 (2007).

[10] Woo, K., Ennew, C. T.: Business‐to‐business relationship quality: An IMP interaction‐based conceptualization and measurement, *European Journal of Marketing*, Vol. 38, pp. 1252–1271 (2004).

[11] Rauyruen, P. and Miller, K. E.: Relationship quality as a predictor of B2B customer loyalty, *Journal of business research*, Vol. 60, No. 1, pp. 21–31 (2007).

[12] 工藤拓, 山本薫, 松本裕治ほか: Conditional Random Fields を用いた日本語形態素解析, 情報処理学会研究報告自然言語処理 (NL), Vol. 2004, No. 47 (2004-NL-161), pp. 89–96 (2004).

[13] Järvelin, K. and Kekäläinen, J.: Cumulated gain-based evaluation of IR techniques, *ACM TOIS*, Vol. 20, No. 4, pp. 422–446 (2002).