一般セッション | 一般セッション | [GS] J-13 AI応用

[1P2-J-13] AI応用: 金融と経済

座長:田部井 靖生(理研AIP)　評者:吉川 友也(千葉工業大学)

2019年6月4日(火) 13:20 〜 15:00 P会場 (1F 展示ホール左奥)

[1P2-J-13-01] インデックス投資が証券市場の価格形成に与える影響の分析

〇松浦 出[1]、和泉 潔[1]、坂地 泰紀[1]、松島 裕康[1]、島田 尚[1] （1. 東京大学）

13:20 〜　13:40

[1P2-J-13-02] 金融機関のテキストデータを活用した景気センチメントの計測

〇近藤 浩史[1]、與五澤 守[1]、成瀬 道紀[1]、森 正和[1] （1. 日本総合研究所）

13:40 〜　14:00

[1P2-J-13-03] 機械学習を用いた地域間の仮想通貨フローの可視化

〇全 珠美[1]、水野 貴之[2,1] （1. 総合研究大学院大学、2. 国立情報学研究所）

14:00 〜　14:20

[1P2-J-13-04] 火災事故が被災企業に及ぼす経済的影響の把握に向けた統計的分析

〇佐藤 遼次[1]、佐藤 一郎[1]、水野 貴之[2] （1. 東京海上日動リスクコンサルティング株式会社、2. 国立情報学研究所）

14:20 〜　14:40

[1P2-J-13-05] 新聞記事からの因果関係を考慮したアナリストレポートの自動要約文生成

〇高嶺 航[1]、和泉 潔[1]、坂地 泰紀[1]、松島 裕康[1]、島田 尚[1]、清水 康弘[2] （1. 東京大学、2. 野村證券株式会社）

14:40 〜　15:00

The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

インデックス投資が証券市場の価格形成に与える影響の分析

Investigating the Effect of Index Investing on Stock Price Formation

松浦 出 *1 Izuru Matsuura　和泉 潔 *1 Kiyoshi Izumi　坂地 泰紀 *1 Hiroki Sakaji　松島 裕康 *1 Hiroyasu Matsushima　島田 尚 *1 Takashi Shimada

*1東京大学 大学院工学系研究科
School of Engineering, the University of Tokyo

In this paper, we modeled stock markets to investigate the effect of index investing on stock price formation. We showed that index investing has little effect on stock price formation in our stock markets model by analyzing results from experiments with various market settings.

1. はじめに

インデックス投資とよばれる投資法がある．投資する資産を，すべての株式に，その時価総額の比で按分して投資するというものである．[Sharpe 64] に始まる一連の研究を理論的背景に持つこの投資法により運用される資産は，現在では投資信託の総運用資産の無視できない割合を占めるに至っている．

インデックス投資では，企業の業績を全く勘案せずに投資が行われる．そのためインデックス投資があまりに大きなシェアを占めた場合，証券市場での価格形成が適正に行われず，有望な企業に資金が集まらない，あるいは投資に値しない企業に資金が集まってしまう可能性が考えられる．

本論文では，インデックス投資が本当にこのような価格形成の問題を引き起こすのか，また引き起こすとすれば，それはどの程度価格形成に影響を与えるのかを検証した．具体的には，証券市場に存在する証券と市場参加者，および価格決定をモデルとして設計し，いくつかのパラメータについてのシミュレーション実験を通して影響を分析した．

2. 証券市場のモデル

本節では，本研究で扱う証券市場のモデルについて述べる．

証券市場には m 種類の証券が存在し，その 1 株あたりのペイオフ $\theta = (\theta_1, \theta_2, \cdots, \theta_m)^T$ は，平均 $\overline{\theta}$，共分散行列 Σ の正規分布 $\mathcal{N}\left(\overline{\theta}, \Sigma\right)$ に従う．また無リスク資産が存在し，その利率は 0 である．時間 $t = 0$ で，すべての証券が 1 単位だけ供給される．

証券市場には n 社のファンドが参加している．うち $n-1$ 社がインデックス投資を行わないファンド（以下ではアクティブファンドと呼ぶ）であり，1 社がインデックス投資を行うファンド（以下インデックスファンドと呼ぶ）である．

すべてのファンドは時間 $t = 0$ で投資し，$t = 1$ に回収する．回収までにポートフォリオを組み替えることはできない．

$n-1$ 社のアクティブファンドはすべて絶対的リスク回避度一定型の効用関数を持つ．j 番目のアクティブファンドは $t = 0$ で θ に関するシグナル $s_j = \theta + \varepsilon_j + \eta$ を受けとる．ここで，θ は次期の証券のペイオフ，ε_j はファンドに固有の誤差，η はすべてのファンドに共通して入りこむ誤差を表し，$\varepsilon_j \sim \mathcal{N}\left(0, \Sigma_\varepsilon\right), \eta \sim \mathcal{N}\left(0, \Sigma_\eta\right), \left(\Sigma_\varepsilon = \sigma_\varepsilon^2 \Sigma, \Sigma_\eta = \sigma_\eta^2 \Sigma\right)$ を仮定する．アクティブファンドは受けとったシグナルをもとに，自身の期待効用を最大化するよう行動する．

連絡先: m2018imatsuura@socsim.org

ただ 1 社存在するインデックスファンドは各証券の時価総額に応じて投資する．インデックスファンドの証券 k への投資額は，証券 k の時価総額がすべての証券の時価総額の和に占める割合に，運用総資産を掛けた額である．

2.1 アクティブファンドの行動

各アクティブファンドは，自身の期待効用を最大化するよう投資する．アクティブファンドの効用関数には絶対的リスク回避度一定型を仮定する．すなわち j 番目のアクティブファンドの効用関数 $u_j(x)$ は，$t = 1$ での運用資産 y に対して，

$$u_j(y) = -\mathrm{e}^{-\rho y} \tag{1}$$

と表せる．$\rho > 0$ は絶対的リスク回避度である．ρ はすべてのアクティブファンドに共通であると仮定する．

この状況では，シグナル s_j を得たアクティブファンド j の投資 x_j は次の最適化問題の解である．

$$\text{maximize} \quad \mathrm{E}\left[u_j\left(\theta'^{\mathrm{T}} x_j\right) \middle| s_j\right] \tag{2}$$

$$\text{subject to} \quad p^{\mathrm{T}} x_j = b_j \tag{3}$$

ただし θ' は m 種類の証券に無リスク資産を加えた $\theta' = (1, \theta_1, \cdots, \theta_m)^T$ のことであり，p は $t = 0$ での無リスク資産を含む証券の市場価格ベクトル $p = (1, p_1, \cdots, p_m)^{\mathrm{T}}$，$b_j$ はアクティブファンド j の運用資産である．

実はこの最適化問題は次の問題と等価である．*1

$$\text{maximize} \quad \theta_j^{\mathrm{T}} x - \frac{\rho}{2} x^{\mathrm{T}} \Sigma_j x \tag{4}$$

$$\text{subject to} \quad p^{\mathrm{T}} x_j = b_j \tag{5}$$

θ_j, Σ_j はそれぞれ s_j を所与とした θ'， Σ' の条件付き平均，条件付き分散である．この解は方程式，

$$\begin{pmatrix} \rho\Sigma_j & p \\ p^{\mathrm{T}} & 0 \end{pmatrix} \begin{pmatrix} x_j \\ \lambda \end{pmatrix} = \begin{pmatrix} \theta_j \\ b_j \end{pmatrix} \tag{6}$$

を解くことで得られる．

2.2 インデックスファンドの行動

インデックスファンドはマーケットポートフォリオに投資する．すなわち，証券価格 $p = (p_1, \cdots, p_m)^{\mathrm{T}}$ と運用資産 b に対して，各証券に $b\,(p_1, \cdots, p_m)^T / \sum_i p_i$ だけ投資する *2．

*1 証明は付録参照

*2 すべての証券の供給を 1 に正規化しているので，時価総額は価格と等しい．そのため時価総額の比が $(p_1, \cdots, p_m)^T / \sum_i p_i$ となる

2.3 価格の決定方法

証券の価格は，すべての証券の超過需要が 0 となる価格で決定される．[*3] 価格 $p \in \mathbb{R}^m$ における証券の超過需要 $D(p) \in \mathbb{R}^m$ とは，$D_j(p) \in \mathbb{R}^m$ を，価格が p であるときの第 j ファンドの需要として，次の式で定義される量である．

$$D(p) = \sum_{j=1}^{n} D_i(p) - 1 \tag{7}$$

超過需要を 0 とする p は，p についての方程式 $D(p) = 0$ をニュートン法で解くことで得られる．

2.4 各証券がどの程度正確に価格付けられているかの指標

インデックスファンドが存在せず，アクティブファンドが受け取るシグナルが，$\theta + \eta$ であるとき，つまり $\sigma_\varepsilon = 0$ の場合を，最も正確な情報が反映された市場と考える．この市場で実現する価格 p_f を完全情報価格と呼ぶことにし，これをベンチマークとする．ある市場がどの程度情報を反映しているかを，その市場で実現する価格 p_p が完全情報価格からどの程度離れているか表す指標

$$d(p_p) = \frac{\|p_p - p_f\|}{\|p_f\|} \tag{8}$$

により評価する．以下ではこの指標を完全情報価格からの乖離度と呼ぶ．

3. シミュレーション実験による検証

3.1 パラメータの決定法

本モデルでは，証券市場はファンドの数 n，証券の種類 m，ペイオフの期待値 $\bar{\theta}$ と共分散行列 Σ，アクティブファンドの絶対的リスク回避度 ρ，ファンドに固有の誤差の大きさ σ_ε，共通の誤差の大きさ σ_η のパラメータで決定される．証券に関するパラメータ $m, \bar{\theta}, \Sigma$ は，東証第 1 部の 1 業種を 1 つの証券と対応させて，2010 年 1 月から 2018 年 9 月までの業種別時価総額の月次データから定める．2010 年 1 月から 2018 年 9 月までの業種毎の時価総額の月次成長率 μ とその共分散行列 Σ_μ をモーメント法により推定する．2018 年 9 月の業種 $i \in \{1, 2, \cdots, 33\}$ の時価総額 S_i を，すべての業種の時価総額の和 $\sum_{i=1}^{33} S_i$ で割ったものを s_i とし，$s = (s_1, \cdots, s_{33})^{\mathrm{T}}$ とする．これらの値を用いて，$\bar{\theta} = \mu * s$, $\Sigma = (ss^{\mathrm{T}}) * \Sigma_\mu$ (ただし $*$ は要素ごとの積) とする．

ρ の値に応じて投資行動がどのように変化するのかを見るために，簡単な例を挙げる．確率 0.8 で賭け金が 2 倍に，確率 0.2 で 0 になるギャンブルを考える．パラメータ ρ の絶対的リスク回避度一定型効用を持つ人が資産 b のうち w だけをこのギャンブルに回すとする．彼のギャンブルへの投資額 w は，X を成功確率 0.8 のベルヌーイ分布に従う確率変数として，

$$\mathrm{E}\left[-\mathrm{e}^{-\rho(1-w+2wX)}\right] \tag{9}$$

を最大化するよう決定される．この関数を最大化する w は，

$$w = \frac{15}{16\rho} \tag{10}$$

である．絶対的リスク回避度一定型の効用関数を持つ人は，その資産の多寡にかかわらずギャンブルへの投資額を決める．[*4] $\rho = 1$ であれば 15/16 を，$\rho = 15$ であれば 1/16 を，$rho = 150$ であれば 1/160 をギャンブルに回すわけである．実験の詳細の節で詳述する通り，本実験では各アクティブファンドの運用資産は 1/160 から 1/16 程度である．前述のギャンブルのような有利な投資案への投資額がこのようなものであることを考えると，絶対的リスク回避度 ρ は 1 から 64 程度に設定するのがよいだろう．

アクティブファンドの推定誤差の大きさ σ_ε^2,σ_η^2 については，どのような値が適切であるか見当がつかないため，$(\sigma_\varepsilon^2, \sigma_\eta^2) \in \{1, 2, 4, 8, \cdots, 64\}^2$ のすべての場合について調べ上げる．

$\sigma_\varepsilon^2, \sigma_\eta^2$ の値について，もう少し詳しく解釈しておく．まずシグナル $s_j = \theta + \varepsilon_j + \eta$ のもとでの θ の条件付き期待値と分散は，

$$\mathrm{E}\left[\theta | s_j\right] = \bar{\theta} + \Sigma\left(\Sigma + \Sigma_\varepsilon + \Sigma_\eta\right)^{-1}\left(s_j - \bar{\theta}\right) \tag{11}$$

$$= \frac{\sigma_\varepsilon^2 + \sigma_\eta^2}{1 + \sigma_\varepsilon^2 + \sigma_\eta^2}\bar{\theta} + \frac{1}{1 + \sigma_\varepsilon^2 + \sigma_\eta^2} s_j \tag{12}$$

$$\mathrm{Var}\left[\theta | s_j\right] = \Sigma - \Sigma\left(\Sigma + \Sigma_\varepsilon + \Sigma_\eta\right)^{-1}\Sigma \tag{13}$$

$$= \frac{\sigma_\varepsilon^2 + \sigma_\eta^2}{1 + \sigma_\varepsilon^2 + \sigma_\eta^2}\Sigma \tag{14}$$

である．各アクティブファンドの受け取るシグナルがどの程度信頼できるかは $\sigma_\varepsilon^2 + \sigma_\eta^2$ のみによって決定される．$\sigma_\varepsilon^2 + \sigma_\eta^2$ が小さければ小さいほど，ファンドは自身が受け取ったシグナルをより信頼する．σ_ε^2 と σ_η^2 個別の値は，ファンド間のシグナルがどの程度ばらつくかにのみ影響する．

3.2 実験の詳細

インデックスファンドの運用資産が市場に占める割合と，完全情報価格からの乖離度の関係を調べるために，次のような実験を行った．$m, \bar{\theta}, \Sigma$ はすべての実験で前述の，東証 1 部の業種別時価総額から定めた値を使う．またファンドの数は $n = 16$ を採用する．各実験では，パラメータ $\rho, \sigma_\varepsilon, \sigma_\eta$ の値を 1 つ選んだ．インデックスファンドの運用資産 C_i を決定し，アクティブファンドの運用資産を $(1 - C_i)/(n - 1)$ で定めた．シード値を 1 から 20 まで変化させ，それぞれのシード値について，この市場で実現する価格 p_p を計算し，同じシード値を用いて完全情報価格 p_f を計算した．結果として得られた p_p の p_f からの乖離度 $d(p_p)$ と，C_i の値との関係を評価した．

3.3 結果と考察

図 1 は，$\rho = 16$ を固定して，$\sigma_\varepsilon, \sigma_\eta$ ごとに横軸に C_i を，縦軸に $d(p_p)$ をプロットしたものである．インデックスファンドの運用資産 C_i の大きさにかかわらず，完全情報価格からの乖離度はほぼ一定のようである．C_i が大きくなるにつれて完全情報価格からの乖離度が線形に増加する傾向があるようには見えるが，この原因はおそらく次のようなものである．

本実験ではアクティブファンドの効用関数に絶対的リスク回避度一定型を仮定したので，アクティブファンドの運用資産の額によらず，アクティブファンドがリスク資産に投資する額は一定である．一方インデックスファンドは運用資産のすべてを

---

*3 この価格決定モデルは経済学で一般均衡理論として知られるものである．詳細は [Jean- Pierre 07]1 章などを参照．

*4 ファンドの効用関数に絶対的リスク回避度一定型を仮定すると，ファンドの運用資産の規模によってその性質が大きく変わってしまう．ファンドが投資家にある一定の性質を持つ金融商品を提供するものであると考えると，この特徴を持つ効用関数を採用するのはあまり適切ではない．当該仮定は計算時間を削減するための技術的なものである．

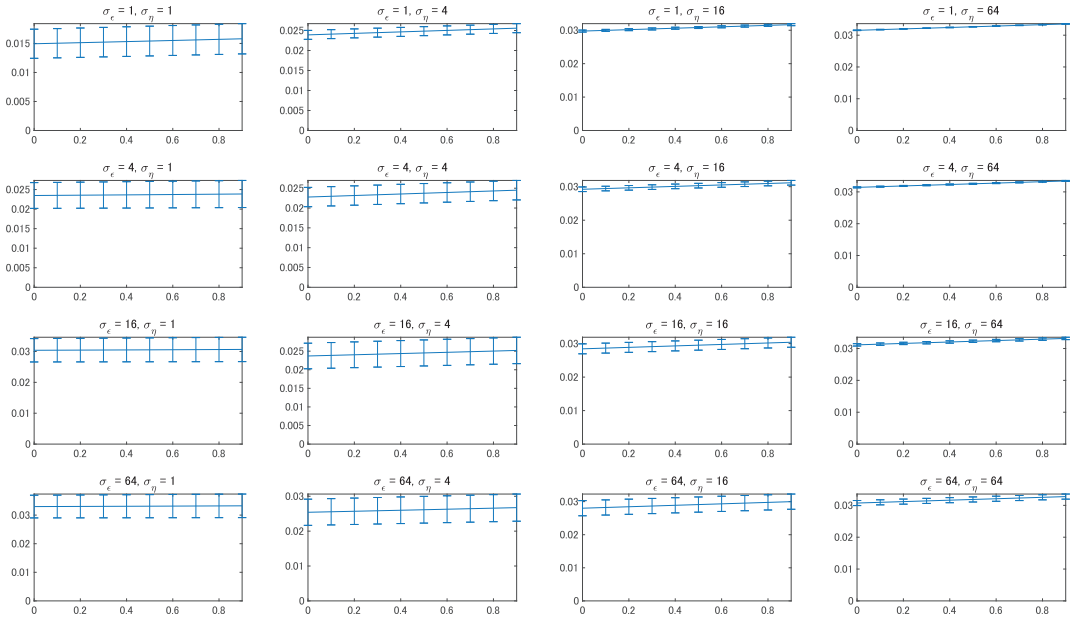

図 1: $\sigma_\varepsilon^2, \sigma_\eta^2$ ごとの，インデックスファンドのシェア（横軸）と完全情報価格からの乖離度（縦軸）との関係

リスク資産に投資するので，C_i の値に応じて市場全体でのリスク資産への投資額が線形に増加する．リスク資産に投資された額が大きいほど各リスク資産の価格は上昇して，$C_i = 0$ の状況で計算された完全情報価格から乖離するはずである．

ρ の値を $\rho = 1, 2, 4, 8, 16, 32, 64$ と変化させて同一の分析をしたところ，$\rho = 16$ のときと同じく，C_i と $d(p_p)$ の間には非常に小さな線形の関係が見られた．

以上のことから，本論文で設定したモデル上では，インデックスファンドが市場の価格形成に与える影響はほとんどないと言えるだろう．

本実験の結果は，本質的に次の 2 つの仮定によるものであると考えられる．

- アクティブファンドのシグナルの正確さ，つまり σ_ε^2 の小ささが，ファンドの運用資産に依存せず一定であると仮定したこと
- アクティブファンドの効用関数に絶対的リスク回避度一定型を仮定したこと．

この 2 点を考慮すると結果が大きく変わりうる．インデックスファンドのシェアが大きくなったとき，前者はアクティブファンドの運用資産の減少によって，分析に使える資源が減り，受け取るシグナルが正確でなくなることを通じて，後者は正確な情報を持ったアクティブファンドがリスク資産への投資額を減少させることを通じて，市場価格と完全情報価格との乖離を大きくすると考えられる．

前者を考慮するには，σ_ε^2 をアクティブファンドの総資産の大きさに反比例するように定めればよい．

後者を考慮するには，アクティブファンドの効用関数に相対的リスク回避度一定型効用関数を仮定すればよい．相対的リスク回避度一定型の効用関数を持つアクティブファンドの意思決定は，本モデルの絶対的リスク回避度一定型効用関数を持つアクティブファンドの意思決定に比べて，計算資源を要すると考えられる．

当面の課題はこの 2 つの仮定を緩和したモデルの作成と実装である．

4. まとめ

インデックス投資が証券市場の価格形成に与える影響を見るために，証券市場のモデルを作成した．またそのモデル上での実験結果から，インデックス投資が価格形成にほとんど影響を与えていないことを示した．

今後はアクティブファンドの効用関数を相対的リスク回避度一定型に差し替え，またアクティブファンドの予測能力が運用資産に依存する構造をモデルに導入し，市場への影響の変化を分析したい．

参考文献

[Jean- Pierre 07] Jean- Pierre Danthine, Donaldson, J. B.: 現代ファイナンス分析 資産価格理論, ときわ総合サービス (2007)

[Sharpe 64] Sharpe, W. F.: Capital asset prices: A theory of market equilibrium under conditions of risk, *The journal of finance*, Vol. 19, No. 3, pp. 425–442 (1964)

A 付録

A1 最適化問題 (2) と最適化問題 (4) の等価性

最適化問題 (2) は，問題 (4) と等価である．これは次のようにして示せる．

$Z = \theta'^T x_j$ とし，Z の s_j の条件付き密度関数を $f_{Z|s_j}(z|s_j)$

とすると，

$$\mathrm{E}\left[u_j\left(\theta'^{\mathrm{T}}x_j\right)\middle|s_j\right] = \mathrm{E}\left[-\exp\left(-\rho Z\right)|s_j\right] \tag{15}$$

$$= -\int \mathrm{e}^{-\rho z} f_{Z|s_j}(z|s_j)dz \tag{16}$$

である．ここで，θ', s_j の両方が正規分布に従うことから，s_j のもとでの θ' の条件付き分布は正規分布である．したがって，その期待値を θ_j，分散を Σ_j とすると，$Z = \theta'^{\mathrm{T}}x_j$ の s_j のもとでの条件付き分布は，期待値 $\bar{z} = \theta_j^{\mathrm{T}}x_j$，分散 $\sigma^2 = x_j^{\mathrm{T}}\Sigma_j x_j$ の正規分布である．これを用いると，

$$\begin{aligned}
&-\int \mathrm{e}^{-\rho z} f_{Z|s_j}(z|s_j)dz \\
&\quad = -\int \mathrm{e}^{-\rho z}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(z-\bar{z})^2}{2\sigma^2}\right)dz \\
&\quad = \int \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{\left(z-\bar{z}+\rho\sigma^2\right)^2}{2\sigma^2}\right)dz \\
&\qquad \times\left(-\exp\left(-\rho\bar{z}+\frac{\rho^2}{2}\sigma^2\right)\right)
\end{aligned}$$

が得られる．この式の最右辺の被積分関数は，期待値 $\bar{z}-\rho\sigma^2$，分散 σ^2 の正規分布の密度関数であるからその積分値は 1 である．ゆえに式 (2) は $-\exp\left(-\rho\bar{z}+\frac{\rho^2}{2}\sigma^2\right)$ と等しい．

以上により式 (2) の最大化は，$-\exp\left(-\rho\bar{z}+\frac{\rho^2}{2}\sigma^2\right)$ の最大化に帰着することがわかった．さらにすべての x_j と x'_j に対して，

$$-\exp\left(-\rho\bar{z}+\frac{\rho^2}{2}\sigma^2\right) > -\exp\left(-\rho\bar{z}'+\frac{\rho^2}{2}\sigma'^2\right) \tag{17}$$

$$\Leftrightarrow \exp\left(-\rho\bar{z}+\frac{\rho^2}{2}\right) < \exp\left(-\rho\bar{z}'+\frac{\rho^2}{2}\sigma'^2\right) \tag{18}$$

$$\Leftrightarrow -\rho\bar{z}+\frac{\rho^2}{2}\sigma^2 < -\rho\bar{z}'+\frac{\rho^2}{2}\sigma'^2 \tag{19}$$

$$\Leftrightarrow \bar{z}-\frac{\rho}{2}\sigma^2 > \bar{z}'-\frac{\rho}{2}\sigma'^2 \tag{20}$$

$$\Leftrightarrow \theta_j^{\mathrm{T}}x_j-\frac{\rho}{2}x_j^{\mathrm{T}}\Sigma_j x_j > \theta_j^{\mathrm{T}}x'_j-\frac{\rho}{2}x'^{\mathrm{T}}_j\Sigma_j x'_j \tag{21}$$

であるから，最適化問題 (2) は別の最適化問題 (4) に帰着する．

金融機関のテキストデータを活用した景気センチメントの計測

Measuring Economic Trends based on Financial Institution Texts

近藤 浩史[*1] Hirofumi Kondo　與五澤 守[*1] Mamoru Yogosawa　成瀬 道紀[*1] Michinori Naruse　森 正和[*1] Masakazu Mori

[*1] 株式会社 日本総合研究所
The Japan Research Institute, Limited

Despite statistics released by the government or central banks have been used to grasp the economic trends, there is a lag in the timing of the survey and publication of the results. Therefore, there have been a lot of research to estimate them ahead of the publication. In this research, we tried to quantify the economic sentiment indicator by analyzing the huge amount of text data created and accumulated in financial institutions. As a result, it was found that our indicator had a high correlation with the Bank of Japan TANKAN, short-term economic survey of enterprises in Japan, and also had a quick reporting nature.

1. はじめに

グローバル化・IT 化により経済情勢の変化速度が早まり，企業は迅速に現状の経済情勢や景気概況を把握する事が重要となっている．従来，経済情勢等を把握するには，政府や中央銀行が公表する統計を活用してきた．一方，このような統計は，調査から公表に一定時間を要するため，統計を補完し，かつ速報性のある新しい指標が必要となってきている．

新しい指標の構築に向け，[経済産業省 2017]では民間企業のデータである SNS や POS 等のビッグデータから AI 技術を活用し，速報性に優れた景気指標の開発に取り組んでいる．

他にも AI 技術を活用し，中央銀行が発行する公的な文書や SNS 等の情報から景況感を示す指数の構築や，経済指標の推定が試みられてきた[饗場 2018][大和証券株式会社 2017][余野 2018]．特に[饗場 2018]では，Twitter から「抽出 AI」を用いて景気に関するツイートを抽出し，得られたツイートに対して「評価 AI」を用いることで景況感を示すセンチメントを算出している．算出した指数は景気ウォッチャー調査[*1] 現状判断 DI（全国：原数値）と高い相関が得られたと報告されている．

本研究では，金融機関が保持するテキストデータを活用して，景況感を反映し，かつ速報性のある景気センチメント（一定期間の景況感を反映した指数）の計測を目的とする．特に本研究は，金融機関の社員が取引先企業との面談を通して作成したテキストデータ（以下，計測対象テキストと記載）を活用して景気センチメントを計測する．

金融機関は日常的に企業の経営者や担当者と接する機会があり，金融機関が作成するテキストデータには，経済活動と関連する記載を含む可能性が高い．また，SNS とは異なり，内容も一定の品質が保たれているため，景気センチメントを計測する元データとして有望と考える．

結果として，本研究で計測した景気センチメントは景況感を示す代表指標である日銀短観と高い相関を示すことが分かった．また，事前に予測できないイベント（地震等）が発生した場合の景況感について，速報性がある可能性を示唆した．

2. 景気センチメント指数の構築

先行研究[饗場 2018]を参考とした．計測対象テキストから景気に関連する文（以下，景気関連文と記載）を抽出し，各文の景況感を数値化して景気センチメントを計測する．本節はそれぞれの実現方法を述べる．

2.1 計測対象テキストと前処理

計測対象テキストは 2006 年 1 月～2018 年 8 月に作成されたテキストである．計測対象テキストには経済環境等に触れた内容も含むが，景気とは全く関連しない内容も多数含む．

計測対象テキストは複数の文書から成り，それぞれの文書を文に分割して使用した．文を単語に分割する際には MeCab[*2] を使用し，辞書として mecab-ipadic-NEologd[佐藤 2017]と独自の金融用語辞書を組み合わせて使用した．

また，意味ある文を構成しないと想定される短文（動詞，名詞，形容詞の合計が 5 単語以下の文）は事前に除去した．このようにして全体で約 5,000 万件の文を得た．以下では，計測対象テキストの前処理済みの文を計測対象文と呼ぶ．

2.2 景気関連文の抽出

景気関連文の抽出では，景気ウォッチャー調査の景気判断理由集を学習データとして活用する．

景気関連文抽出モデルは入力文が景気ウォッチャー調査の景気判断理由集に含まれる文（以下，調査文と記載）か，計測対象文かを分類する文章分類モデルである（図 1）．

学習済みモデルに計測対象文を入力し，調査文と分類された文は，景気判断理由集に含まれる文に類似する文，すなわち景気関連文と見なせる．つまり，計測対象文を入力したにもかかわらず，モデルが調査文と誤判定した文を収集することで，景気関連文を抽出する．

文章分類モデルとして，以下に示す 5 つのモデルを作成・比較した．

連絡先：日本総合研究所，東京都品川区東五反田 2 丁目 18 番 1 号大崎フォレストビルディング，03-6833-0900（代）

*1: 内閣府 景気ウォッチャー調査 (https://www5.cao.go.jp/keizai3/watcher/watcher_menu.html)
*2: MeCab : Yet Another Part-of-Speech and Morphological Analyzer (https://taku910.github.io/mecab)

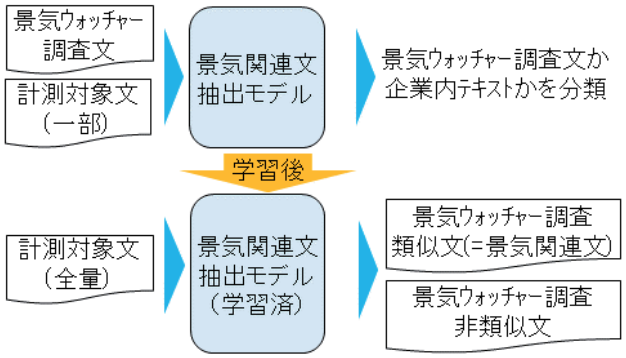

図 1：景気関連文抽出モデルの概念図

- 【TF-IDF/LR モデル】訓練時に使用した文に含まれる単語の TF-IDF を特徴量として，ロジスティック回帰モデルで文章を分類するモデル．TF-IDF の特徴量として，単語の出現頻度が 5 以下の単語は無視した．
- 【CNN/NN モデル】先行研究[Kim 2014]をベースとしたモデル．
- 【双方向 LSTM（BiLSTM）/NN モデル】双方向 LSTM（以降，BiLSTM と記載）とニューラルネットワークを組み合わせたモデル．
- 【SWEM/LR モデル】[Shen 2018]に SWEM-concat と記載されたモデル．SWEM-concat の特徴量を文の特徴量として，ロジスティック回帰で文章を分類するモデル．
- 【アンサンブルモデル】景気関連文の抽出精度向上を目的とし，上記 4 つのモデルをアンサンブルしたモデル．各モデルから出力された調査文らしさのスコアが，予め決めた閾値よりも大きい場合に，文を抽出する．

各モデルは調査文 (2013 年 1 月～2018 年 8 月)および計測対象文を元に学習させた．学習ではそれぞれ 5 万文（計 10 万文）をランダムに選択して用いた．評価にはそれぞれ 6 千文（計 1.2 万文）をランダムに選択して文章分類モデルの性能を評価した．訓練済みの単語の分散表現が必要な場合は，計測対象テキストから学習させた 200 次元の word2vec（Skip-gram）モデル[Mikolov 2013]を作成・利用した．

表 1 は文章分類モデルの性能評価の結果である．どのモデルも高精度で調査文と計測対象文を分類できる．

表1：文章分類モデルの性能

	精度	再現率	F 値
TF-IDF/LR	0.986	0.989	0.987
CNN/LR	0.990	0.993	0.992
BiLSTM/NN	0.987	0.983	0.985
SWEM/LR	0.979	0.983	0.981

2.3 文の景況感の数値化

先行研究[山本 2016]と同様に，深層学習を使用した回帰モデルを構築した．学習済みモデルに文を入力すると，入力文が内包する景況感が数値化されて出力される．本研究では景気ウォッチャー調査の景気判断理由集の文（2013 年 1 月～2018 年 8 月）から学習データ（153,913 件）および評価データ（17,104 件）をランダムに選択・使用して，モデルを構築・評価した．

テストデータに対する平均二乗誤差は 0.309 となり，先行研究[山本 2016]と同等の結果を得た．ここでも訓練済みの単語の分散表現が必要となるため，前述と同様の word2vec モデルを使用した．

表 2:文章分類モデルごとの相関係数

モデル名	相関係数/ t 値(N=51)	四半期ごと 計測対象文数(千件)
TF-IDF/LR	0.843 / 11.0	11.0
CNN/NN	0.831 / 10.5	8.4
BiLSTM/NN	0.834 / 10.6	6.7
SWEM/LR	0.868 / 12.2	14.8
アンサンブル（閾値 2）	0.823 / 10.2	7.2

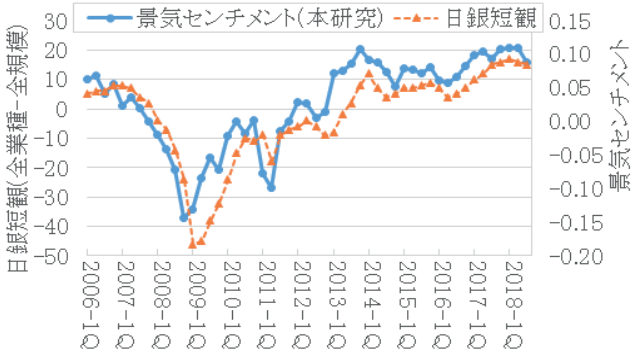

図 2：日銀短観と景気センチメント(景気関連文のみ)の時系列

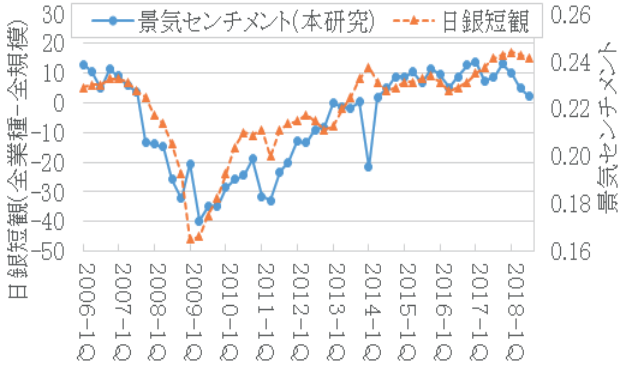

図 3：日銀短観と景気センチメント(計測対象全文)の時系列

2.4 景気センチメントの計測

各文の景況感を示す数値を一定期間ごとに集計し，景気センチメントとして計測する．本研究では，予め決めた期間ごとに，各文に付与された景況感の数値の単純平均を計算する．

例えば，四半期単位の景気センチメントを計測する場合，3 ヶ月ごと，当該期間の文につき，景況感の数値の平均を算出する．

3. 評価

3.1 日銀短観との相関分析

計測した景気センチメントが世の中の景況感を示す値であるかを検証するため，日銀短観-業況判断-最近（全産業-全規模）（以下，日銀短観と記載）との相関を分析した．なお，日銀短観の公表頻度は四半期である．また，日銀短観の第四四半期の指数は 12 月中旬に公表されるが，景気センチメントの第四四半期の計測にあたっては 12 月末までを集計範囲としている．

表 2 に景気関連文の抽出モデルを変えた場合の，景気センチメントと日銀短観の相関係数を記載した．どのモデルを使用しても，計測した景気センチメントは日銀短観と高い相関関係

(係数 0.8 以上)を示すことが分かる. 特に SWEM とロジスティック回帰を使用する抽出手法が最も良かった.

景気関連文の抽出精度向上を狙って閾値を 2 とした場合のアンサンブルモデルは係数が低下する. これは, 景気関連文を絞り込み過ぎると, 計測対象となる文数が減少し, 一部の文が示す景況感に左右されるためと考えられる. 集計対象となる文数が少ない CNN/NN や BiLSTM/NN においても, 同様の傾向が見られる.

今回使用した計測対象テキストに対しては, 景気関連文と似ている文章を幅広に取得する手法を用いることで, 良い景気センチメントを算出できる.

3.2 日銀短観と景気センチメントの時系列推移

図 2 に日銀短観と景気関連文から計測した景気センチメントの推移を図示した. なお, 文の抽出モデルは SWEM/LR を使用して計測した. 図 2 より, 景気関連文から計測した景気センチメントは日銀短観におよそ追従することが読み取れる. 特にリーマンショック(2008 年)や, 東日本大震災(2011 年)のイベント時には連動している様子が顕著に読み取れる.

一方, 図 3 は, 景気関連文の抽出効果を示すために, 計測対象文の全文から景気センチメントを計測した場合の時系列推移である. 図 3 を見ると, 計測対象文の全文から計測した景気センチメントも全体的な動きは日銀短観におよそ追従するが, 異なる動きをしている点もある. 特に 2014 年第一四半期では両者が大きく反対に動いている点が挙げられる.

図 2 と図 3 から, 日銀短観との追従という観点においては, 景気関連文を抽出することで, 良い景気センチメントを計測できることが分かる. これは, 計測対象文から景気に無関係な文, すなわち, 計測のノイズとなる文を除去できたためと考える.

3.3 指数の速報性

図 2 によると計測した景気センチメントは日銀短観を先行していると読み取れる部分もある.

例として東日本大震災直後に公表された 2011 年第一四半期の指数に着目する. 東日本大震災の発生により, 日本の景況感は悪化したはずであるが, 日銀短観は震災発生の直後に悪化の方向に動いていない. これは, 震災が 2011 年 3 月 11 日に発生した一方, 日銀短観の企業へのヒアリング期間が 2011 年 2 月 24 日～3 月 31 日であるため, 日銀短観が震災の影響を完全に反映しきれていない可能性があるためと推測できる. 一方, 計測対象テキストは震災発生後も作成されており, 3 月 11 日以降の景況感を強く反映した値を算出できたと考える.

このように, 計測した景気センチメントは震災などの事前に予期できないイベントが発生した場合の景況感として, 速報性のある数値を算出している可能性がある.

参考までに, 月次で公表されている景気ウォッチャー調査現状判断 DI(全国:原数値)との相関を分析すると, こちらも高い相関(係数:0.838, t 値:18.8, N=152)を示す(図 4). 景況感を示す月次数値としての活用も可能と考える.

なお, [饗庭 2018]にて算出された SNS×AI 景況感指数と景気ウォッチャー調査 DI との相関係数は 0.79 と報告されており, 単純な比較はできないものの, 本研究の景気センチメントのほうが高い相関を示している.

4. まとめと今後の方針

本研究では金融機関のテキストデータから景況感を示す景気センチメントを計測した. テキストデータから景気に関連する文のみを抽出し, 景況感を数値化することで, 日銀短観と高い相関を持つ景気センチメントが計測できることを示した. また, 震災などの予期しないイベントが発生した場合に, 速報性のある景況感を計測できる可能性も示した.

今後は, 他景気指標との関係を分析することや, 得られた景況感の数値を用いた将来予測の可能性も検討したい. また, 日銀短観との相関についても継続的に分析し, 本研究の有効性をモニタリングする予定である.

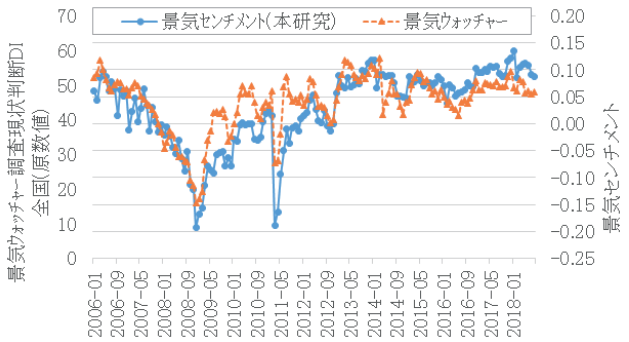

図4:景気ウォッチャー調査と景気センチメントの時系列推移

5. 免責事項

本稿は著者らの見解を示すものであり, 所属機関の公式見解を示すものではありません.

参考文献

[経済産業省 2017] 経済産業省 : 平成 28 年度 IoT 推進のための新産業モデル創出基盤整備事業(ビッグデータを活用した新指標開発事業) 報告書, 経済産業省大臣官房調査統計グループ調査分析支援室委託調査, 2019.

[饗場 2018] 饗場 行洋, 山本 裕樹 : データサイエンスと新しい金融工学, 財界観測(2018 春号) , 2018.

[大和証券株式会社 2017] 大和証券株式会社 : 株式会社大和総研 :「大和地域 AI(地域愛)インデックス」の公表について, プレスリリース, https://www.dir.co.jp/release/2017/20170713_012138.html (2018/2/4 アクセス), 2017 .

[余野 2018] 余野 京登, 和泉 潔, 坂地 泰紀 : 金融レポート, およびマクロ経済指数による日銀センチメント指数の構築, 第 32 回人工知能学会全国大会 , 2018.

[山本 2016] 山本 裕樹, 松尾 豊 : 景気ウォッチャー調査の深層学習を用いた金融レポートの指数化, 第 30 回人工知能学会全国大会, 2016.

[佐藤 2017] 佐藤 敏紀, 橋本 泰一, 奥村 学 : 単語分かち書き辞書 mecab-ipadic-NEologd の実装と情報検索における効果的な使用方法の検討, 言語処理学会第 23 回年次大会, 2017.

[Kim 2014] Yoon Kim : Convolutional Neural Networks for Sentence Classification, EMNLP2014, 2014.

[Shen 2018] Dinghan Shen, Guoyin Wang, Wenlin Wang, Martin Renqiang Min, Qinliang Su, Yizhe Zhang, Chunyuan Li, Ricardo Henao and Lawrence Carin : Baseline Needs More Love: On Simple Word-Embedding-Based Models and Associated Pooling Mechanisms, ACL2018, 2018.

[Mikolov 2013] Mikolov Tomas, Sutskever Ilya, Chen Kai , Corrado Greg and Dean Jeff : Distributed Representations of Words and Phrases and their Compositionality, NIPS2013, 2013.

The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

機械学習を用いた地域間の仮想通貨フローの可視化

Visualization of Inter-Regional Flows in the Virtual Currency using Machine Learning

全 珠美[*1]
Joomi Jun

水野 貴之[*1,*2]
Takayuki Mizuno

[*1] 総合研究大学院大学
SOKENDAI

[*2] 国立情報学研究所
National Institute of informatics

Human activity creates a specific pattern in 24 hours. The patterns can be influenced by the time zone they live in. Therefore, we can classify their time zone by their activity pattern. We have built a bitcoin-time zone classifier using XGBoost, a machine learning approach. We have classified the time zone of the specific bitcoin addresses used in the Ponzi event and visualized the flow.

1. はじめに

激しい価値当落やその財貨としての信頼性に疑問を挙げるにも関わらず，仮想通貨への人々の興味や投資今でも続いている．ビットコインの場合，2017 年の最高値以来，値段も取引量も下がっているが，いまだ仮想通貨市場で最大の出来高を維持している[CM 19]．我々は代表的な仮想通貨であるビットコインにおける各ユーザーの取引の日中パターンを機械学習により分類することで，ユーザーの活動地域を推定し，サイバー空間で行われている仮想通貨の流れを実空間に可視化する．

ビットコインは実空間に関して強い匿名性を持っていると思われるが，ユーザーやビットコインの流れを追跡することは不可能ではない．アドレスのクラスタリングからユーザーを特定する手法[Reid 12]，ビットコインのノード間のメッセージからノードの IP を推定する手法[Kaminsky 11, Juhasz 16]などが提案されている．このようにビットコインの匿名性を解消しユーザーの空間情報を明らかにする研究は行われている．

サイバー空間でのユーザーの空間情報を推定する手法はソーシャルメディアでもよく研究されている．例えば，Twitter におけるユーザーの投稿パターンから地域情報を推定することができる[Mahmud 14]．

本稿では，はじめに，仮想通貨コミュニティウェブサイト Bitcointalk.org で活動するビットコインユーザーの活動関連データ（活動地域，投稿履歴，取引履歴）を分類学習機の学習データとして活用し，取引の地域を分類する分類器を作成する．作成した分類器を用いて，活動地域が不明な十分に取引数があるユーザーに関して，取引パターンから実空間における活動地域を推定する．

次に，推定された活動地域を用いて，特定のビットコインの実空間における流れを分析する．我々は 2015 年から 2016 年に掛けて Bitcointalk.org に掲載され行われた特定の Ponzi イベントに注目した．ビットコイン市場の Ponzi イベントは，ユーザーがあるアドレスにビットコインを送ると，イベント主催者は集まったビットコインで投資や事業をおこない，ビットコインを送ったユーザーには何パーセントかの利益が上乗せされて返還される．一般的に，イベント参加者が増えると，初期参加者が利益を得る構造があり，乗り遅れまいと多くのユーザーが殺到する．そして，しばしば，後期参加者には，ビットコインの返還がなされない．我々は，被害者のいる Ponzi イベントにおける実空間でのビットコインの流れを可視化し，どの地域で活動するユーザーが利益を，また，被害を受けているのか明らかにする．

連絡先：全 珠美，総合研究大学院大学（情報学専攻），東京都千代田区一ツ橋 2－1－2，joomi@nii.ac.jp

2. データセット

ユーザーの活動パターンを分析するため二つのデータセットを用意した．一つ目はビットコインと関連して地域情報が分かるユーザーのデータである．我々は仮想通貨コミュニティウェブサイト Bitcointalk.org における 2009 年 12 月から 2018 年 9 月までの，ビットコインに関する 3,251,067 件の投稿記事と投稿日時，そして，その間の全 891,795 人の投稿者に紐づく，ユーザー名，ユーザーレベル(投稿頻度等)，ユーザーの地域情報（タイムゾーン）を収集した．本サイトの地域情報は GMT に初期設定されているため，地域情報が GMT になっているユーザーは解析から除外した．

二つ目はビットコインの取引履歴データである．公開されているビットコイン取引の情報の中から，ビットコイントランザクション ID，直接やり取りしたビットコインアドレス，ビットコインの量，直前・直後のトランザクション ID，そして，Bitcointalk.org でユーザーが公開したビットコインアドレスの取引データを，www.walletexplorer.com を利用して集めた．ビットコインアドレスは，ウォレット形式にクラスタリングされている．また，主要なウォレットは，「取引所」や「マイニング」など取引種別の TAG が付けられている．

3. ビットコインユーザーの地域推定

人間は一日中必ず休みの時間が発生するが，その時間帯は地域（タイムゾーン）によって変わる．日本人とアメリカ人の日中活動を GMT に変換しパターン化すると，その形が異なる．つまり，ユーザーの活動パターンが地域（タイムゾーン）の影響を受けるため，そのパターンで地域を推定することができる．

我々はまず，Bitcointalk.org で公開されているユーザーのビットコインアドレスデータから，サイトにおける投稿のパターンとビットコイン取引パターンに注目した．Fig. 1 は投稿と取引のパターンを表しており，二つのパターンが類似していることが読み取れる．

次に，Bitcointalk.org のユーザー投稿パターンを用いてビットコイン取引が行われる地域を推定する分類器を構築する．ユーザーの投稿パターンデータと地域情報を学習データセットとしてた．パターンを十分に学習させるために，学習データは，日中 5 時間以上の時間帯で投稿を行ったユーザーに限定した．地域の分類は，ヨーロッパ EU（GMT+1~3），アジア ASIA（GMT

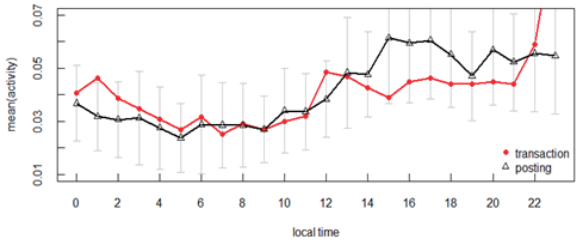

Fig. 1 : ユーザーの投稿パターンとビットコイン取引パターン

	投稿データ	取引データ
Accuracy	0.888	0.91
95% CI	(0.86, 0.912)	(0.871, 0.939)
No information rate	0.333	0.333
Kappa	0.832	0.865

Table 1 : ユーザー地域分類学習の結果

+7~9)とそれ以外の地域 OTH の三つに分類した. ヨーロッパとアジアのユーザーは今回使ったデータの 89%を占めていた. アップサンプリングをおこなうことにより, サンプルサイズの偏りをなくした.

本稿ではユーザーの活動パターンから地域を推定する分類器の構築に, 機械学習の一種である XGBoost[Chen 16]を採用した. XGBoost は Random Forest アルゴリズムを基盤とし, Gradient Boosting を結合した学習法である. 速度の遅い Gradient Boosting を並列処理で行っているため学習や分類の信頼性と速度が優れている. 我々は, 学習データをトレニングデータセット(80%)とテストデータセット(20%)に別け, XGBoost で学習した結果が Table 1 である. 地域推定を 90%に近い精度でおこなえていることが分かる.

4. イベントにおけるビットコインの地域間の流れ

構築した地域(タイムゾーン)分類器を利用し, 特定の Ponzi イベントにおける地域間のビットコインの流れを分析する. はじめに, イベント主催者のウォレットを中心に, 直前(送った)・直後(貰った)取引先の地域を分類する. Fig. 2 の地域別取引回数から, EU ユーザーを対象としたイベントであることが読み取れる.

次に, イベント主催者のウォレットに送る前の3取引先と, 主催者のウォレットから受け取った後の3取引先を調査し, 地域間の流れを観測する. Fig. 3 は, 前・後三つまでの取引先の地域を分類し, 各地域間におけるビットコインの出来高(BTC 量)を表している. 同地域間の取引が多いことが分かる. つまり, 匿名性を持つビットコイン市場であっても, 物理的距離の近いユーザー同士が主に取引している.

物理的距離が近いユーザー間の取引が多い理由の一つは, ビットコイン取引ネットワークにおけるハブの存在である. 取引所やサービス系会社(ATM や Payment)のウォレットがハブの役割をしており, このバブの活動パターンは, しばしば地域的特徴を強く持っている. 各地域のユーザーは, その地域のこのようなハブを好んで利用する傾向があり, その結果が, 地域内の取引が多い要因になっている. 今回, 分析を行った Ponzi イベントの前後の取引で利用された TAG 付がなされたウォレットの中でも, 取引所が 24.9%, サービス系会社が 25.3%を占めていた.

5. おわりに

本研究では機械学習手法 XGBoost を利用し, ビットコインコミュニティにおけるユーザーの投稿パターンとビットコイン取引のパターンからユーザーの地域を推定する分類器を作成した. こ

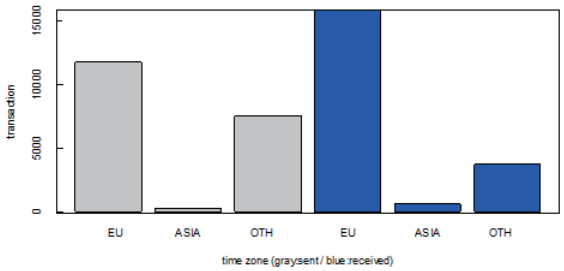

Fig. 2 : イベント参加者の地域別取引回数

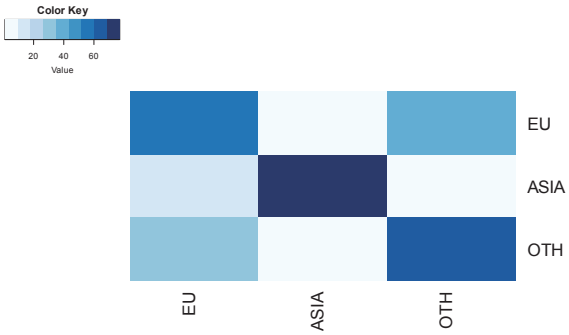

Fig. 3 : 地域別ビットコインの流れ(%)
(X : Target / Y : Source)

の分類器を用いて特定の Ponzi イベントに参加したビットコインユーザーの地域を推定し, ビットコインの実空間での流れを調査した. このイベントは, EU ユーザー向けに開催されており, 各地域内で主に取引されていたビットコインが, ユーザーがこのイベントに参加することで, EU に流れ込み, その後, それらのコインは EU 内で流通していることが分かった. つまり, このイベントで各国に被害者がいるとすると, 被害者のビットコインは主に EU 域内で流通していると言える.

ビットコイン以来, 多種多様な仮想通貨が生み出され, その利用範囲は拡大している. 仮想通貨は, 主にサイバー空間で取引されるが, その取引は実空間での社会生活や経済活動と密接に関係している. この関係性を理解するためには, 実空間に射影した流れや利用の特徴を把握する必要がある. 本研究のような, 仮想通貨の流れを実空間に可視化する努力がこれからも必要になる.

参考文献

[CM 19] https://coinmarketcap.com/coins/(Accessed 2019 Feb)

[Reid 12] Reid,F.; Harrigan, M. : An Analysis of Anonymity in the Bitcoin system, Security and Privacy in Social Networks, 197-223, 2012

[Kaminsky 11] Kaminsky D. : Black Ops of TCP/IP, Presentation, Black Hat & Chaos Communication Camp, 2011

[Juhasz 16] Peter L. Juhasz, Jozsef Steger, Daniel Kondor, Gabor Vattay. : A Bayesian Approach to identify Bitcoin Users. PLos ONE, 2016

[Mahmud 14] Jalal Mahmud, Jefferey Nichols, Clemens Drews : Home Location Identification of Twitter Users, ACM Transactions on Intelligent Systems and Technology,2014

[Chen 16] Tianqi Chen, Carlos Guestrin. : XGBoost: A Scalable Tree Boosting System, In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016

The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

火災事故が被災企業に及ぼす経済的影響の把握に向けた統計的分析

Statistical analysis aimed at assessing the economic impact of fire accidents on damaged companies

佐藤 遼次[*1]　佐藤 一郎[*1]　水野 貴之[*2]
Ryoji Sato　Ichiro Sato　Takayuki Mizuno

[*1] 東京海上日動リスクコンサルティング株式会社　企業財産本部
Property Risk Engineering Department, Tokio Marine & Nichido Risk Consulting Co., Ltd.

[*2] 国立情報学研究所
National Institute of Informatics

We aim to clarify the trends of fire impacts on corporate finance and stock markets through statistical and machine learning analysis on data sets of companies suffered from fire accident. In this paper, we confirmed that the effect on corporate finance become larger on the fire accidents having certain characteristics in the newspaper articles. We also confirmed that the trend of the impact on corporate finance can be predicted by considering the type of industries and certain characteristics in newspaper articles in a complex manner. We will also report the result of the analysis about trend of impact on stock market in this conference.

1. はじめに

企業が火災事故により被災した場合，生産設備等の資産の損失，事業継続性の損失などの直接的な影響に始まり，株価の下落という形で株式市場へも影響を及ぼし得る．実際の個々の火災事故に対する直接的な影響は，上場企業であれば，公開されている財務指標から確認することが（例えば[鈴木 2012]のように）可能である．また株価への影響についても，過去の株価データにアクセスすることで，個々の被災企業における影響を確認することは容易である．

しかしながら，企業における火災事故全体を考えた場合，企業財務や株式市場に有意に影響を及ぼしているのか，またそれはどういった特徴を持つ火災事故の場合なのか，といった疑問について明らかにしようとした研究は，著者らの知る限り存在しない．こうした疑問について明らかにすることは，防火・防災活動に取り組む当事者企業や，そうした活動を支援する様々な機関・企業にとって有益と考えられる．

そこで本研究では，過去の火災事故事例から構築したデータセットに対する統計的および機械学習を用いた分析を通じて，火災事故が企業財務や株式市場へ及ぼす経済的影響の傾向を明らかにすることを目的とする．本稿ではまず，新聞データベース及び企業情報のデータベースから，多数の火災事故事例を整理したデータセットを作成する．次に，それらの情報を企業の財務指標データベースと照合・分析することで，火災事故による直接的な被害の傾向について明らかにする．更に，その結果を株価のデータベースとも照合することで，直接的被害の傾向との対比から，株式市場への影響の傾向についても明らかにすることを目指す．

2. 手法

2.1 火災事故データセットの作成

まず，新聞データベース及び企業情報のデータベースから，企業の火災事故に関するデータセットを作成した．新聞データベースとしては，（株）日本経済新聞社の記事データベースから，火災事故に関係する記事データを抽出したものを使用した．抽出したデータの諸元を表 1 に示す．

次に，各記事に付与されている株式コードを検索キーとして，火災に関する記事を企業単位で抽出した．抽出対象とした企業は，表 2 に示す 7 業種，計 657 の上場企業である．なお，企業の業種分類は，日本経済新聞の日経業種分類（[日本経済新聞社]）に基づいている．また，持株会社設立等により過去に株式コードが変更されている企業については，変更前の株式コードも極力検索キーに含めるよう補完して抽出を行った．結果として，計 71 の企業に対して，火災事故の発生を報じる記事を抽出した．

また，一つの企業において複数件の火災事故が報じられている場合には，それらを区別して整理することで，計 156 件の火災事故を特定した（表 2）．以上を通じて，各火災事故に関する新聞記事の情報を整理することで，火災事故及び被災企業の情報をまとめたデータセットを作成した．

2.2 財務指標データセットの準備

財務指標に関するデータベースには，トムソン・ロイターの QA Direct（1980～2016 年）を使用した．このデータベースには，企業が有価証券報告書や四半期報告書において公開している財務指標の値が纏められている．本研究ではこのデータベースから，以下に記す 2 つの財務指標を利用した．

1 つ目は，特別損失（EXTRAORDINARY_CHARGE-PRETAX）である．有価証券報告書等における決算報告とは，企業会計原則などの各種基準に基づき作成されており，その中で，災害による損失は特別損失として計上することが定められている．そのため，火災事故により発生する固定資産の滅失・損失，損壊した資産の点検費，撤去費用などを，「火災による損失」等の勘定科目で特別損失に計上することが一般的となっている．従って，特別損失の値を利用することで，火災事故による被害を定量的に捉えることができると考えた．

連絡先：佐藤遼次，東京海上日動リスクコンサルティング（株）企業財産本部，東京都千代田区大手町 1-5-1，03-5288-6234, ryoji.sato@tokiorisk.co.jp

表 1 ：抽出した新聞記事データの諸元

対象媒体	日本経済新聞，日経産業新聞，日経MJ，日経地方経済面
抽出条件	テーマ分類：火災（＃W50202）コード付与※1，且つ，企業が主題の記事※2
対象期間	1975 年 1 月 1 日～2018 年 6 月 10 日
抽出件数	4,865 件

※1 データベース側で自動付与されるコード．火災に関連する単語の知識辞書，及び前後の単語との共起などから判定される．
※2 データベース側で自動付与される属性．企業名の登場回数や登場位置などの条件を考慮して判定される．

表 2 :抽出した火災事故の内訳

業種番号	業種名	企業数	火災が報じられている企業数	火災事故件数
1	自動車	76	12	27
2	ゴム	20	1	4
3	鉄鋼	47	12	49
4	化学	207	30	51
5	機械	234	4	8
6	繊維	49	6	10
7	パルプ・紙	24	6	7
-	合計	657	71	156

但し，ここで取り扱う特別損失の値には，火災事故以外の災害損失，或いは固定資産売却損，有価証券の売却損など，火災事故とは無関係な損失が一定程度含まれる可能性がある点に留意が必要である．また，滅失した固定資産に火災保険が掛けられている場合には，保険金の確定を待った上で，保険金で賄われない分の損失だけが特別損失として計上される場合がある．従って，「火災による損失」等として計上されている金額は，火災事故による損失の全てを含んでいない可能性がある点に留意が必要である．

2 つ目の財務指標は，売上高（NET_SALES_OR_REVENUES）である．これは，特別損失の値を絶対値として扱うのではなく，企業の規模の違いを考慮に入れて分析を行うために利用する．

以上 2 つの財務指標に関する決算期ごとの時系列データを，先に抽出した火災事故の発生年月日と照合することで，火災事故の発生前後における財務指標のデータセットを作成した．

2.3 企業財務への影響を測る指標の設定

火災事故が及ぼす企業財務への影響を評価するための指標として，2.2 までで抽出した財務指標データより，以下の指標を設定した．

$$log_{10}\left(\frac{loss}{sales}\right) \quad (1)$$

ここで，*loss* は火災事故により被災した日を含む通期の特別損失（円）を，*sales* は火災事故により被災した期の 1 つ前の期における通期の売上高を表している．そして，特別損失を売上高で割り，対数スケールに変換することで，式(1)の通り指標を構築した．また，元々の財務指標データベース側で特別損失または売上高の値が欠損しており，指標が計算できない火災事故 35 件を除外し，計 121 件の火災事故に対して式(1)の値を求めた（特別損失が 0 円の場合は，1,000,000 円に置き換えて計算した）．

最後に，式(1)に対して以下の通り規格化処理を行った．まず，2.1 にて抽出した 71 の企業を対象に，式(1)を財務データの全期間に適用して，決算期別の平均値を求め，時系列で描画したグラフを図 1 に示す．この図より，式(1)の値は，業種によらず，時系列上で概ね共通した傾向を示すことが確認できる．これは即ち，式(1)に含まれる火災事故以外の要素（火災事故以外の特別損失）は，個別企業ごとの経営状況などの内部環境よりも，景気等の外部環境によって一律に受ける影響が強く作用していることを示している．そこで，異なる時期の火災事故を適切に比較するため，各火災事故に対する式(1)の値から，該当の決算期における，式(1)の値の全企業平均値(μ)を差し引くことで，新たに指標 式(1')を以下の通り定義した．

$$log_{10}\left(\frac{loss}{sales}\right) - \mu \quad (1')$$

指標 式(1')は，正または負の値をとり，値が大きいほど，火災事故による企業財務への影響が大きいことを示している．この指標を用いて，火災事故による企業財務への影響を分析することとした（以下，「指標」とは式(1')を指すこととする）．

3. 分析と考察

3.1 火災事故の有無による企業財務への影響の評価

まずは，単純に火災事故の有無によって，指標の値の傾向に有意な違いが現れるのかについて検討した．即ち，“火災事故で被災した決算期の特別損失”を用いて計算した指標の値（「火災有り指標」と呼ぶ）と，“火災事故で被災していない直近の決算期の特別損失”を用いて計算した指標の値（「火災無し指標」と呼ぶ）を比較して，統計的に有意な違いが見られるのかについて確認した．

火災有り指標，火災無し指標それぞれの値を，火災事故ごとにプロットした散布図を図 2 に示す．これらの 2 標本に対して，対応のある/対応の無い両側 t 検定による平均値の検定，Kolmogorov-Smirnov 検定による分布の検定を実施したところ，いずれも有意な差は認められなかった（表 3）．

ここで，火災有り指標が扱っている火災事故の中には，大小様々な規模の火災が含まれている．従って，これらの火災事故を一律に抽出しただけでは，火災事故が無い決算期と比べて，企業財務への影響が有意に確認できるわけではないことが示された．

3.2 企業財務に影響を及ぼす火災事故が持つ特徴の分析

3.1 の結果を踏まえ，どのような特徴を持つ火災事故であれば，企業財務への影響が有意に大きくなるのかについて検討した．ここで，2.1 で整理した各火災事故に関する新聞記事から得られる情報を用いて，以下のような特徴を設定した．

- 特徴1：決算への影響に関する単語の有無
  …決算や株価に関する単語が新聞記事中に登場すること
- 特徴 2：操業への影響に関する単語の有無
  …操業や生産に関する単語が新聞記事中に登場すること
- 特徴 3：記事本文の総文字数
  …ある火災事故に関する一連の新聞記事における，記事本文の総文字数が一定以上であること

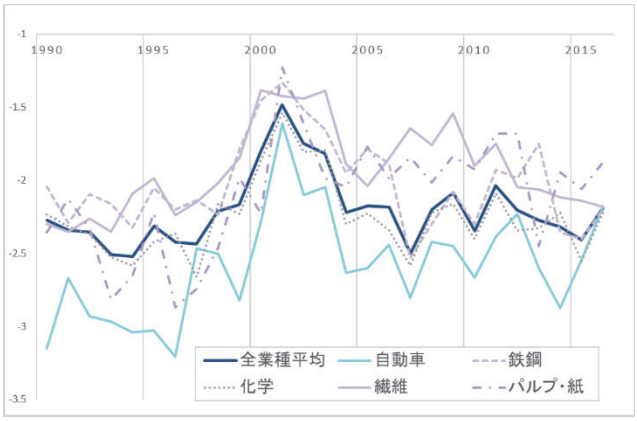

図 1：指標 式(1)の年別平均値

※ 1990 年以前は値の欠損が多いため，また，業種「ゴム」「機械」は企業数自体が少ないため，図から省略.

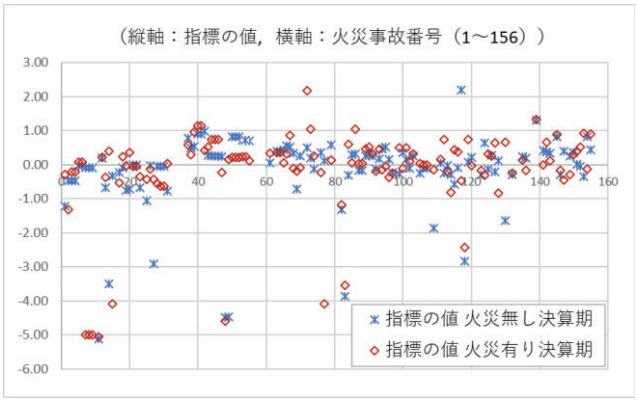

図 2：火災事故の有無に応じた指標 式(1')の値の分布

表 3：火災事故の有り・無しにおける
指標 式(1')の平均と分布の検定

	N	平均	SD	t 値※	D
火災無し_全件	121	-0.14	1.09	0.26 (n.s.)	0.058 (n.s)
火災有り_全件	121	-0.17	1.30		

n.s .有意差なし, *p < .05, **p < .01

（※t 値は対応のある t 検定における値）

以上の 1~3 の特徴に該当する火災事故を判定するため，以下の要領で具体的な基準を設定した.

まず，特徴 1・特徴 2 に関する判定を行うために，日経新聞社によって各記事に自動付与されている「keyword」という情報を利用することとした. このkeyword には，記事の文中から切り出した単語，またはその単語を正式名称に変換したものが格納されている. 次に，企業財務への影響が大きい火災事故にはどのような keyword が格納されているのかを把握するため，Web 上で閲覧できる範囲の有価証券報告書から，“実際に火災による損失が特別損失に計上されている火災事故”を 25 件まで特定した上で，それらの火災に関する新聞記事に含まれている keyword を抽出・整理することで，特徴 1・特徴 2 を定義付ける単語リストを作成した（表 3）. また，特徴 3 については，1 つ 1 つの新聞記事の本文に書かれている文章の文字数を，火災事故ごとに合算することで，総文字数を求めた. その上で，“総文字数 1,500 文字以上”を閾値として設定した.

表 3：特徴 1・特徴 2 を定義する単語リスト

単語リスト_特徴 1（決算への影響に関するもの）	連結決算 / 収益見通し / 純利益 / 決算 / 売上高 / 営業利益 / 減益 / 損失 / 連結 / 特別損失 / 経常利益 / 経常益 / 企業業績 / 株主総会 / 経常増益 / 配当 / 収益 / 利益 / 業績 / 株価 / 連結営業利益 / 連結売上高 / 経常減益 / 予想 / 営業減益
単語リスト_特徴 2（操業への影響に関するもの）	再開 / 操業再開 / 復旧/操業停止 / 減産 / 稼働 / 生産 / 停止 / 操業 / 生産計画 / 生産ライン / 生産動向 / 生産調整 / 生産量 / 生産中止 / 見通し / 回復 / 生産見通し / 生産拠点 / 生産能力 / 委託生産 / 再稼働 / 要請 / 委託 / 生産委託 / 供給停止 / 生産体制 / 工場再開 / 生産開始 / 操業開始 / 設備復旧 / 代替生産 / 長期化 / 生産再開 / 資材調達

以上の通り定義した特徴 1~3 に該当する/該当しない火災事故ごとに指標 式(1')の値を分け，対応の無い片側 t 検定，Kolmogorov-Smirnov 検定をそれぞれ実施した. 結果，有意水準 5%とした場合，いずれの特徴についても，平均値が有意に大きくなること，分布についても有意に異なることが確認された（表 4）.

以上より，企業財務に影響を及ぼす火災事故であるかどうかを判定するためには，記事中に決算や操業に影響することを表す単語が登場しているか，長い文章の記事で繰り返し報道されているか，といった特徴が有効であることが確認された.

3.3 機械学習を用いた企業財務への影響の傾向分析

3.2 より，新聞記事から得られる特定の特徴に着目することで，企業財務に影響を及ぼす火災事故を一定程度予測できる可能性が示唆された. そこで，こうした特徴を説明変数に，2.3 で設定した指標の値を被説明変数として回帰分析を行うことで，企業財務への影響を定量的に予測することを試みた.

回帰分析の手法にはランダムフォレストを用い，実装には Python のモジュールの 1 つである scikit-learn を用いた [Pedregosa 2011]. 説明変数には，3.2 で述べた特徴 1~3 に加え，各企業が属する業種（7 分類）を設定した. モデルの作成に当たっては，まず訓練データ:テストデータ = 7:3 に分割した上で，訓練データに対して 3 分割交差検証によるグリッドサーチを行い，パラメータ（決定木の深さ，バギングに用いる決定木の個数）を決定した. ここで，精度の評価には本来テストデータのみを用いるべきであるが，本研究ではデータ数が限られることから，訓練データ・テストデータを合わせた全データに対してモデルを適用した結果で精度を評価することとした.

モデルを適用した予測結果を図 3 に示す. 実データで外れ値と言える極端に大きな値についてはフィットしないものの，一定程度値の傾向を再現できていることがわかる. この精度を評価するため，実データ及び予測値に対してスピアマンの順位相関係数を求めたところ，相関係数=0.44 となり，有意水準 5%, 1%でそれぞれ有意であることを確認した. 即ち，本モデルにより，実データとの順位相関を棄却することなく，指標 (1')の値を予測できていることが確認された.

また，この予測モデルにおける特徴量の重要度を可視化した結果を図 4 に示す. この図から言えることとして，まず特徴 1~3 については，特徴 2（操業への影響に関する単語），特徴 1（決算への影響に関する単語），といった説明変数が予測に貢献

表 4：特徴 1~3 における指標 式(1')の値の検定結果

	N	平均	SD	t 値	D
特徴 1	22	0.32	0.43	3.60**	0.78**
特徴 2	50	0.05	0.94	1.72*	0.31**
特徴 3	22	0.16	0.44	2.42**	0.78**

n.s .有意差なし, *p<.05, **p<.01

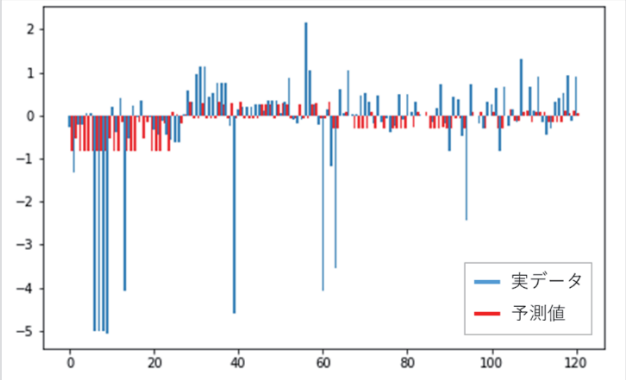

図 3：指標 式(1')の実データと予測値の対比

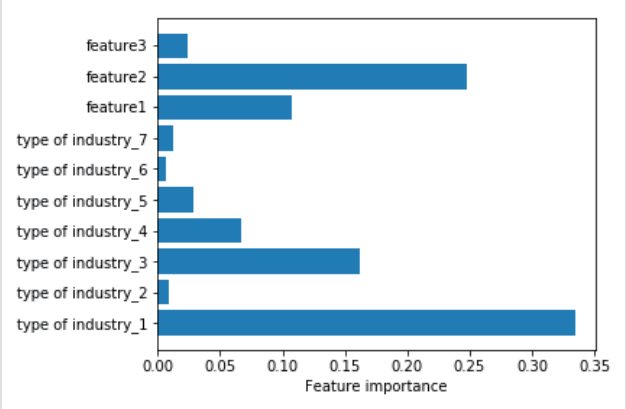

図 4：予測モデルにおける特徴量の重要度
(type of industry の対応については表 2 を参照)

していることがわかる. 3.2 の表 4 では, 特徴 2 は最も平均値が低かったものの, 該当するデータ数が 50 件と多いことから, 他の特徴に比べて, 決定木の分離に多く活用されたものと考えられる. また業種について見ると, 業種 1(自動車)の重要度が特に高くなっている. ここで, 2.3 の図 1 を改めて見ると, 自動車は他の業種に比べて指標の値が全体的に小さくなっており, この特徴によって被説明変数を効果的に分離できたものと考えられる. 実際に, 業種ごとの売上高としては自動車業が最も大きくなっており, 火災で被災したとしても, 売上高対比で見た影響は大きくなりにくい, という特徴が予測に貢献していると言える.

最後に, 以上 3.1~3.3 の結果の総括として, 各特徴を用いて火災事故を絞込んだ場合, およびランダムフォレストによる予測結果の上位に絞り込んだ場合における, 指標値の分布を図 5 に示す. この図より, 各特徴によって企業財務への影響が大きい火災事故を一定程度絞り込めること, また, それらの特徴を業種と併せて機械学習に供することで, 影響の大きい火災事故をより精度良く予測できていることがわかる. これは言い換えれば, 新聞記事による報道の特徴や, 被災企業の業種といった属性を複合的に考慮することが, 火災事故による企業財務への影響を予測する上で重要であることを示している.

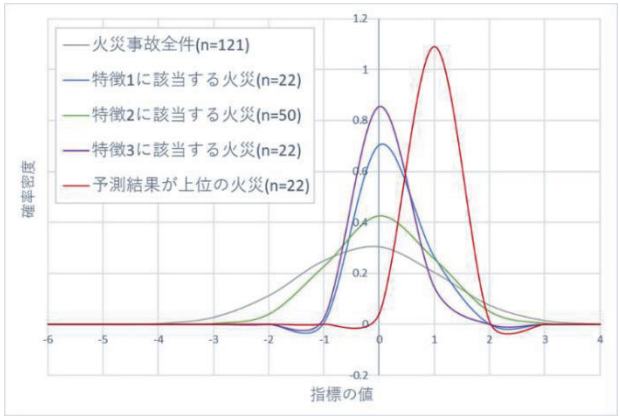

図 5：特徴 1~3 に該当する火災事故，および，機械学習で予測した指標値の上位 22 件における指標値の分布

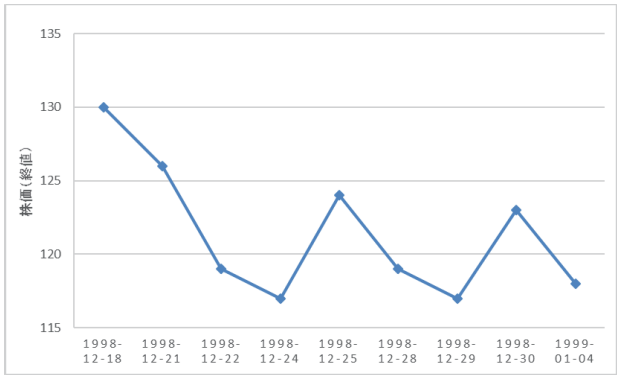

図 6：某企業の火災事故前後における株価（終値）の推移
（火災発生日：1998-12-27）

4. 結論と今後の展望

本研究では, 新聞記事データから特定した過去の火災事故事例を用いて, 火災による企業財務への影響について分析した. 結果として, 単純な火災事故の有無だけでは有意な傾向の違いは確認できないものの, 新聞記事での報道に一定の特徴を持つ火災に限定した場合, 企業財務への影響が有意に大きくなることを確認した. また, 業種や新聞記事に関する特徴を複合的に考慮することにより, 企業財務への影響の傾向を一定程度予測可能であることを確認した.

今後は, 火災事故による株式市場への影響についても分析を行う. 今回特定した火災事故事例のうち, 財務への影響が大きい火災を 1 つ取り上げ, 火災前後における株価(終値)の推移をプロットしたものを図 6 に示す. 実際に, 火災発生直後の取引日における株価の下落が確認できる. こうした傾向について, 企業財務への影響の大きさとの比較を交えて分析を行い, 結果を報告する予定である.

参考文献

[鈴木 2012] 鈴木 拓人:化学工場の爆発火災事故の増加とその影響について, NKSJ-RM レポート, 2012.

[日本経済新聞社] 日経業種分類(2018-11-30 閲覧)
https://www.nikkei.com/markets/company/search/gyoshu/

[Pedregosa 2011] Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python, Journal of Machine Learning Research, 12:2825–2830, 2011.

The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

新聞記事からの因果関係を考慮したアナリストレポートの自動要約文生成

Automatic Summarization of Analyst Reports Based on Causal Relationships from News Articles

高嶺 航 *1 Wataru Takamine
和泉 潔 *1 Kiyoshi Izumi
坂地 泰紀 *1 Hiroki Sakaji
松島 裕康 *1 Hiroyasu Matsushima
島田尚 *1 Takashi Shimada
清水 康弘 *2 Yasuhiro Shimizu

*1東京大学大学院工学系研究科 School of Engineering, The University of Tokyo
*2野村證券株式会社 Nomura Securities Co., Ltd.

In this paper, we focused on the causal relationships in both of news articles and analyst reports. We proposed a novel approach for summarizing analyst reports automatically based on the causal relationships extracted from both text data. As a first step toward summarization of analyst reports adequately, we analyzed the validity of the method in extracting causal relationships which can be evaluated from the analyst reports. As a result, the proposed method could extract basis information of analyst's opinions from analyst reports with some accuracy, and we could confirm the styles of analysts in expression of opinions and bases.

1. はじめに

近年，投資家に対する投資判断の支援を行う技術の必要性が高まってきており，投資判断材料の一つであるアナリストレポートの活用に注目が集まってる．アナリストレポートには，証券市場調査・分析の専門家である証券アナリストが企業の経営状態や収益力などを調査した結果がまとめられており，企業の業績や株価に対する証券アナリストの予想と根拠が示されている．記述されている予想の根拠としては，その企業の取り組む事業の近況・財務状況（企業のファンダメンタルズ），事業に影響を与える経済・政治・社会状況（マクロ経済のファンダメンタルズ）などの外部要因についても言及されている．このように，高度な専門知識をもつアナリストによる詳細なレポートは，株価の変動要因にもなりうる [1] ため，彼らの企業の業績や株価に対する予想やその裏付けとなる根拠を投資判断の材料として活用することは有用性が高いと思われる．

しかしながら，アナリストレポートの発行の多くは決算発表の時期に集中し，膨大なレポートの全てを熟読するのは難しく，レポートの内容を十分に把握できない可能性がある．

この問題に対して，近年，自然言語処理やテキストマイニング技術の進展により，膨大な量のアナリストレポートから重要な要点のみを自動で抽出・要約する技術のニーズが高まっており，研究事例も報告されている [2][3]．このように，投資判断材料に必要な情報を要約することができれば，レポートを読む負担が減り，時間の制約がある中でもレポートの内容の要点を把握することができる．しかしながら，これら [2][3] の要約技術は，テキストに記述されている事象の背景にある因果関係を考慮していない．そのため，生成された要約文に，投資判断材料となりうる証券アナリストの予想の根拠が盛り込まれていない場合が想定される．これに対して，[4] では文の因果関係の構造に注目し，原因表現を取り出す手法を提案している．

このように，アナリストレポートの活用として自動的に重要な箇所を要約・抽出，あるいは検索する技術が研究されている．しかしながら，2 つの異なる媒体 (つまり，アナリストレポートとそれ以外のテキスト情報) から一つの要約文を生成する手法はまだ確立されていない．この手法の確立により，アナリストレポート中で業績・株価予想の根拠として言及される情報の特徴を捉えるだけではなく，新聞記事などの媒体からその根拠の背景についての情報を補うことで，より説明できる情報を含んだ要約文の生成が期待できる．さらに，これが可能になれば，要約の過程で抽出する証券アナリストの株価・業績予想につながる根拠や背景としての経済情報を検索できるようになり，証券アナリストのレポート作成支援としても期待できる．

そこで本研究では，因果関係を考慮しながら別の媒体から補填的に情報を抽出し，要約文を自動生成する手法を提案し，その評価を行う．本稿では，その提案手法実現のために，アナリストレポートから根拠情報を抽出する手法の妥当性について実験を行った．

連絡先: 高嶺航, 東京大学大学院工学系研究科技術経営戦略学専攻, 113–8656 東京都文京区本郷 7–3–1 工学部 8 号館 530 室, m2018wtakamine@socsim.org

2. 提案手法

本章では，テキストデータから因果関係表現を抽出し，要約文を自動生成する手法について述べる．2.1 節では本手法の概観，2.2 節ではアナリストレポートおよび新聞記事からの因果関係表現抽出の概要，そして 2.3 節では，要約文における根拠の背景情報の獲得に用いた表現類似度計算手法の概要を述べる．

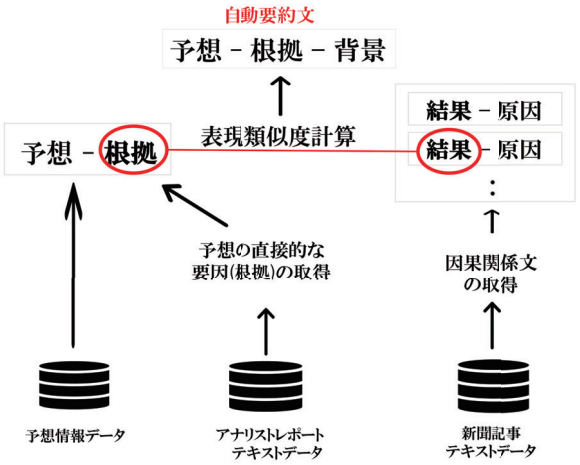

図 1: 提案手法の概説

2.1 要約文生成手法の概説

本節では要約文生成手法の概要を述べる．全体の流れを図 1 に示す．まず，各アナリストレポートにおける証券アナリストの企業業績・株価の予想の情報を獲得する．次に，アナリストレポート本文中に出現する因果関係の構造を抽出し，その中でも結果表現に株価・業績予想が含まれる因果関係を獲得する．獲得した因果関係の原因表現を証券アナリストによる企業業績・株価の予想の根拠情報として獲得する．ただし，この時，証券アナリストの予想情報とその予想の根拠情報は同文章内に出現するものと仮定している．つまり本手法では図 2 のように文章横断的に出現する因果関係表現は抽出しないこととする．また同様に，新聞記事からも因果関係の構造を抽出し，因果関係の結果表現のうち，獲得した証券アナリストの予想の根拠情報と表現が類似する文章を探索する．そして，類似性の高い因果関係における原因表現を根拠情報の背景情報として獲得する．このようにして獲得した，(1) 証券アナリストの企業業績・株価予想，(2) 予想の根拠（直接的な要因），(3) 根拠の背景をまとめ，アナリストレポートの要約文を自動生成する．

本自動要約手法の実装例として，Web サーバー上のシステムとして実装したものの動作画面を図 3 に示す．銘柄，期間を入力すると，すでに生成済みの要約文のうちから入力情報に合致する要約文を出力する．要約文の構成は，一文目に証券アナリストレポートの業績・株価予想，二文目に証券アナリストの予想の根拠，三文目にその根拠の背景を想定している．

文書 1 : 業績予想を下方修正(結果表現)

文書 2 : ○○を織り込んだ。(原因表現)

図 2: 文章横断的に出現する根拠情報の例

図 3: 想定している提案手法を用いたシステムの動作画面

2.2 因果関係抽出手法

アナリストレポート・日経新聞記事から酒井ら [4] の手法を用いて因果関係表現を抽出する．この手法では，因果関係表現を特徴付ける手がかり表現と，手がかり表現に係る節の中で共通して頻繁に出現する共通頻出表現を定義する．最初に少数の手がかり表現と共通頻出表現を与えることで，互いに係り受け関係にある新たな共通頻出表現と手がかり表現が連鎖的に獲得される．この手法を用いる場合，アナリストレポートにおいては特にアナリストの予想を示す文の部分と，その予想の根拠を示す文の部分を分離して抽出する．前者を予想部，後者を根拠部と呼ぶ．

アナリストレポート中から抽出した予想部と根拠部の例を図 4 に示す．この文章の場合，「主力の制御事業の順調な拡大を」を根拠部，「主因に」が手がかり表現，「目標株価を上方修正」が予想部となる．酒井ら [1] は，アナリストレポートからアナリストの予想と根拠情報の抽出を行なっているが，アナリスト予想根拠文の抽出方法として，共通頻出表現の数を用いてアナリストの予想根拠文かどうかを判定している．本手法では，予想の直接的な要因を根拠情報と定義しており，結果表現に業績予想が含まれる因果関係の原因表現を証券アナリストによる企業業績・株価予想の根拠情報として抽出する．

主力の制御事業の順調な受注拡大を

(根拠部)

主因に、目標株価を上方修正

(手がかり表現)　　(予想部)

図 4: アナリストレポートから抽出した予想部と根拠部の例

2.3 表現類似計算手法

本節では，アナリストレポートの根拠の背景情報を新聞記事から獲得するために用いた，二つの文章の表現類似度を計算する手法について述べる．本研究では，表現類似度 s を以下の式のように話題性 t，文の表層 w，極性の一致度 p，文脈の類似性 c の 4 つの構成要素として捉える．

$$s = t \cdot w \cdot p \cdot c \tag{1}$$

- 話題性 : トピックモデル (LDA[5]) による単語の分散表現を用いた文章の話題の類似度を算出
- 文の表層 : Word2vec(Skip-gram[6][7]) による単語の分散表現を用いた文章の表層的な類似度を算出
- 極性の一致度 : 金融極性辞書 [8] を用いた単語の極性を計算し，文章間の極性がどれだけ一致するかを判定
- 文脈の類似性 : アナリストレポートの根拠情報と新聞記事の結果表現の類似度を算出するだけではなく，新聞記事の原因表現との類似度も算出．より根拠情報の文意に沿った文章を抽出する．

LDA と Word2vec を用いて算出された二つのベクトル表現を用いることで，比較する二つの文書の話題性と表層的な類似度を算出する．ベクトル間類似度はコサイン類似度を用いた．A，B はそれぞれ文章 $\vec{A}$，$\vec{B}$ は，それぞれ文書 A，B 内にある名詞・動詞・形容詞の分散表現の相加平均を求めて算出した文書ベクトルである．

$$\text{cosine similarity} = \cos\theta = \frac{\vec{A} \cdot \vec{B}}{\|\vec{A}\|\|\vec{B}\|} \tag{2}$$

表 1: 実験に用いた手がかり表現の概要

手がかり表現の数	手がかり表現の例
109	織り込んで，見込んで，をきっかけに，背景に，考慮し，踏まえ など

3. 実験

提案手法では，おおよそ同一文章内で因果関係表現が出現し，予想とその根拠情報が獲得できるという仮説に基づいて，酒井ら [4] の因果関係抽出手法を使用している．本実験では，対象としているアナリストレポートにおいて，同一文章内で予想の根拠情報が獲得できた件数の割合を示し，根拠情報抽出をする手法の妥当性を検証する．実験には，表 1 に示す手がかり表現を用いた．

実験データには，2011 年から 2016 年までの間に発行された 7927 件のアナリストレポートのうち文章内で因果関係表現が抽出できた 7716 件を用いた．アナリストレポートから抽出した因果表現を含む文章の中から結果表現の部分に「目標株価」および「業績予想」に関する記述がある場合，その原因表現を予想の根拠情報として抽出している．なお，因果関係を抽出するにあたって，本実験では形態素解析器としては Mecab を用い，係り受け解析器としては Cabocha[9] を用いた．

4. 実験結果と考察

評価方法に関しては，文章内で因果関係抽出ができたレポートの件数に対する証券アナリストの予想の根拠情報の抽出ができたレポートの件数の割合を Precision(精度) とした．目標株価に対する根拠情報，業績予想に対する根拠情報，そしてどちらか一方に対する根拠情報が抽出できた割合の 3 項目を算出した．

実験結果を表 2 に示す．検証した全項目で 5 割を下回る精度となり，因果関係ができた抽出した文のうち，結果表現にて目標株価と業績予想のいずれかに言及しているアナリストレポートの件数は，4 割程度であった．必ずしも予想に対して直接的な表現を使用している訳ではないことが分かる．この要因として次の 3 点が挙げられる．

1. 証券アナリストの予想に表記揺れがある（例：野村予想を上方修正，利益予想を引き上げる）
2. 明確な根拠表現を回避する
3. 文章横断的な根拠表現が抽出できない

このうち，1 の予想情報の表記揺れについて検討する．具体的には結果表現に含まれる記述として「目標株価」,「業績予想」に加え,「利益予想」,「野村予想」,「収益」,「売上高」等，計 13 個のフレーズがある場合，予想の根拠情報として抽出を行なった．目標株価および業績予想の根拠情報の抽出割合における表記揺れの考慮の結果を表 2 に示す．

表記揺れを考慮することによって 8 割程度まで精度を向上することができた．予想情報の言及に関して，証券アナリストは複数の言い回しをしており，その表記揺れを考慮に入れて根拠情報を抽出する必要があることが分かった．本結果より，今回実験を行なったアナリストレポートの 8 割程度が予想の根拠情報を同一文章内にて言及しており、本手法における因果関係表現抽出手法の有用性が示すことができた．

表 2: 各予想に対する精度（Precision）

	Precision
目標株価のみ	0.20
業績予想のみ	0.33
目標株価あるいは業績予想	**0.44**

表 3: 表記揺れを考慮に入れた結果（Precision）

	Precision
表記揺れを考慮しない場合	0.44
表記揺れを考慮した場合	**0.83**

5. まとめ

本研究では，新聞記事からの情報を活用し，文章内に出現する因果関係表現と文章間の表現類似性に着目したアナリストレポートの自動要約手法を提案した．その提案手法実現のために，アナリストレポートから根拠情報を抽出する手法の妥当性について実験を行った．表記揺れを考慮することによって，8 割程度のアナリストレポートで同一文章内に証券アナリストの予想とその根拠情報が出現していることが分かり，本論文において用いた因果関係抽出手法の有用性が示された．また，実験結果を通じて証券アナリストのレポート内の書きぶりに関しても考察を行った．今後の課題として，要約文の評価データセットの作成，2 節で紹介した表現類似度計算の精度向上に関する手法の考案，因果関係抽出の精度向上に寄与する手がかり表現の語義曖昧性解消手法の考案などが考えられる．

参考文献

[1] 酒井浩之, 柴田宏樹, 平松賢士, 坂地泰紀. アナリストレポートからのアナリスト予想根拠情報の抽出. 人工知能学会第 17 回金融情報学研究会, pp. 25–30, 2016.

[2] Jahna Otterbacher, Güneş Erkan, and Dragomir R Radev. Using random walks for question-focused sentence retrieval. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pp. 915–922. Association for Computational Linguistics, 2005.

[3] Katja Filippova, Mihai Surdeanu, Massimiliano Ciaramita, and Hugo Zaragoza. Company-oriented extractive summarization of financial news. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 246–254. Association for Computational Linguistics, 2009.

[4] Hiroyuki SAKAI and Shigeru MASUYAMA. Cause information extraction from financial articles concerning business performance. *IEICE Transactions on Information and Systems*, Vol. E91.D, No. 4, pp. 959–968, 2008.

[5] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, Vol. 3, No. Jan, pp. 993–1022, 2003.

[6] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[7] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pp. 3111–3119, 2013.

[8] 伊藤友貴, 坪内孝太, 山下達雄, 和泉潔. テキスト情報から生成された極性辞書を用いた市場動向分析. 人工知能学会全国大会論文集 2017 年度人工知能学会全国大会 (第 31 回) 論文集, pp. 2D21–2D21. 一般社団法人 人工知能学会, 2017.

[9] 工藤拓, 松本裕治ほか. チャンキングの段階適用による日本語係り受け解析. 情報処理学会論文誌, Vol. 43, No. 6, pp. 1834–1842, 2002.