様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　２９　年　　６　月　２０　日現在

機関番号：　２４４０２
研究種目：　基盤研究(B)（一般）
研究期間：　2011～2015
課題番号：　２３３３０１０８

研究課題名（和文）テキストマイニングと高頻度データ解析による金融バブル生成・崩壊過程のミクロ分析

研究課題名（英文）Analysis of the generating and collapsing process of financial bubbles using high frequency stock market data and web-mined text data

研究代表者

高田　輝子（TAKADA, Teruko）

大阪市立大学・大学院経営学研究科・准教授

研究者番号：３０３４７５０４

交付決定額（研究期間全体）：（直接経費）　　14,600,000円

研究成果の概要（和文）：金融バブルの生成・崩壊メカニズム解明に役立つ新事実の発見を目指し、ウェブ上の株式掲示板などから収集したテキストデータと長期高頻度株式時系列データの双方を対象として、これらにLDAトピック推定法やノンパラメトリック適応的確率密度推計ベースの諸方法などの、データ入力のみから効率的情報抽出を実現する推計手法を適用した。その結果、株式数値統計で示された株式市場状態とテキストデータで示された投資家の興味内容やその変動パターンとの間の対応関係や、株価トレンド転換予測に有用な知見を含む、いくつかの新しい事実を明らかにすることができた。

研究成果の概要（英文）：Aiming at finding new empirical facts useful for clarifying the mechanism of financial bubble generation and collapse, this study targets web-mined message board text data and very long high frequency stock market data. Applied methods to those data are efficient and fully data adaptive such as LDA topic model and nonparametric adaptive kernel density estimation based methods. This study revealed several new empirical facts including those useful for forecasting trend reversals and relationship between stock market status indicated by numerical stock market data and the content of investors’ attention captured by web-mined text data.

研究分野：　Computational finance

キーワード：　金融バブル　テキストマイニング　高頻度データ　投資家行動

１．研究開始当初の背景

金融バブルは実際の経済に深刻な影響をもたらすにも関わらず、その発生・崩壊メカニズムはほとんど解明されていない。これは、金融バブル分析の本質的な困難さによるものである。金融バブルが滅多に発生しない事象であるために、分析のために利用可能な情報自体が少ない。これに加えて、突然の大きな変化である、という特性は、通常の統計・ファイナンス手法では扱いづらいという問題がある。

この問題に挑戦するためには、「対象データ自体の増大」と「対象データからの情報抽出の効率化」の二方向から取り組む必要がある。しかし、特に研究開始当初においては、テキストマイニングをはじめとする、Computational intelligence 手法のファイナンス分野への適用例は世界的にも非常に少なく、大規模金融データを用いた解析例も少なかった。こうした分野の応用が進んできた現在においても、本格的な応用研究はまだ限られている。

研究開始当初想定の計画から、一点、変更の必要が発生した。当初は、Yahoo!Finance 株式掲示板から DOW30 指数構成銘柄の数十年分の長期テキストデータをウェブマイニングにより収集することを想定していたが、研究開始後まもなく、提供元のシステム変更によりデータ提供量が大幅に削減され、書込数の多い銘柄については半年程度分のデータしか取得できなくなるというトラブルに見舞われ、長期データ収集の断念を余儀なくされた。そのため、利用可能データで実行可能な形で研究目的を目指す方向に、研究計画を修正することとなった。

２．研究の目的

本研究の目的は、金融バブルの生成・崩壊過程についての、新しい事実を発見することである。これは、金融バブルの原因因子の特定や、因果関係、生成・崩壊メカニズムの解明に寄与するのみならず、財市場のブームや群集心理にまで応用が期待され、極めて研究意義が高いものである。

そのために、利用可能な統計情報量の最大化に、二方向から迫る。まず「対象データ自体の増大」については、これまでファイナンス分野での活用が少なかったテキストデータを株式市場に関係するウェブ情報を対象に、ウェブ上から自動収集する。ウェブ検索頻度統計やウェブ上の株式掲示板書き込み情報を用いる。また、世界最大規模を誇るニューヨーク証券取引所逐次データをはじめとする各種統計を長期間分収集し、全期間を対象にした分析を行う。これにより、利用可能な情報量自体を増やすことができる。

次に、「対象データからの情報抽出の効率化」のために、高度な統計情報技術を用いることにより、収集したテキストデータ及び高頻度株式数値データからの情報抽出量を増大し、分析に利用可能な統計情報量の最大化を実現する。

テキストデータの主力として想定していた株式掲示板情報については、投資家行動と株式市場動向の関係についての、テキスト情報から得られる新事実発見へと目的を修正し、長期的な金融バブル変動についての事実発見については、長期高頻度株式数値統計により行うこととした。

３．研究の方法

Yahoo!Finance 株式掲示板からのデータ収集システム構築および収集作業と並行しながら、以下の２つのアプローチに分けて、研究を行った。

(1) テキストデータを中心にした投資家行動と株式市場状態についての事実発見：単語出現頻度分析及びトピック分析結果と株式市場状態との関係分析

① データ：
テキストデータ収集先は、検索頻度統計については Google trend、株式掲示板データについては Yahoo!Finance 掲示板の書き込み内容データである。どちらのデータについても、自動収集システムを構築し、収集した。

② 単語出現頻度析：
収集したデータを品詞分解し、株式統計により示される株式市場の局面ごとにタグ付けを行い、パターン抽出や予測を行った。

③ トピック分析：
入力文書データから自動的にトピック(話題や分野などの大ざっぱな意味集合)を推定するトピックモデルと呼ばれる機械学習の手法を用いる。具体的には、Latent Dirichlet Allocation (LDA)と呼ばれる、入力文書データ情報のみから、どのように文書内情報のトピック分類を行う手法を用いた。

(2) 長期高頻度株式統計の包括的解析：

長期高頻度株式時系列データを用いて、株価上昇トレンド期、下落トレンド期、方向転換点前後の株式市場の振舞について、以下の手法を用いて解析した。

① 適応的カーネル確率密度推定法
株価のような裾厚(大きな変化が頻繁に起きる)な確率分布に強い、ノンパラメトリック適応的確率密度推計法を用いて、確率密

度形状の特徴分析を行った。

② 局所相互情報量推定法
これまでの相互情報量の推定では、２変数間のの平均値が利用されてきていたが、Takada(2012)が推計方法の高精度化を実現したため、入力データのみから２変数間の非線形な相関構造全体のパターンを推定することが可能になった。これを用いて、全データ点における局所相互情報量を裾部分に至るまで高精度に推計し、各種変数間の関数関係の解析を行った。

４．研究成果

本研究の各アプローチの主な成果は、以下の通りである。

(1) テキストデータを中心にした投資家行動と株式市場状態についての事実発見：出現単語頻度及びトピック分析結果と株式市場状態との関係分析

① 検索頻度統計解析による、投資家の期待度と株式市場動向の関係分析：
検索行動は、興味を向けた対象についての情報収集行為であるため、検索頻度統計解析は、投資家の興味の強さやその変動パターンを解析するのに非常に適している。Google trend における主要株価指数名の検索頻度と当該株価指数の変動パターンを解析し、投資家行動のいくつかのアノマリーや、暴落を引き起こす因子が内生的である可能性など、いくつかの新しい事実を発見した。成果の一部は高田・佐藤(2012)で発表された。

② 株式掲示板書込内容の意味･内容分析：
Yahoo!Finance 株式掲示板の LDA トピック分析により、株式掲示板書込データから意味が理解できるトピックを取り出すことができた。従来の株式掲示板書込内容を利用したテキスト分析は、書込内容の意味がわからない状態で株価予測への有効性を論じる観点からのものがほとんどであり、意味が理解できる形でトピック内容について分析した例は、まだないと思われる。佐藤・井上・高田(2015)で発表された。

③ 株式掲示板から抽出したトピックと日次株式統計が示す株式市場状態との関係分析：
ファイナンス分野におけるテキスト分析では辞書を利用したものが多いが、ニュースではなく、投資家心理を観察する上で、様々な問題があることを確認した。一方、本研究が採用している LDA トピック分析のアプローチは、文書データ入力のみからトピック抽出を行うアプローチであるため、そのような問題は発生しない。これを利用し、株式掲示板上のトピックとボラティリティーや取引量などの株式市場動向との間の、いくつかの新しい事実を発見した。佐藤・高田(2017)で発表された。

④ 株式掲示板から抽出したトピックと高頻度株式統計が示す株式市場状態との関係分析：
株式掲示板内容に関する様々な変数とリターン、ボラティリティ、取引量をはじめとする株式市場状態を示す様々な変数との間の包括的な非線形関係を、局所相互情報量推定により可視化した。また、株式掲示板トピック内容と株式市場動向との間に有意な関係が発生する最短時間間隔を明らかにした。これは、株式掲示板分析の高頻度化の目安を示すだけでなく、投資家の期待･不安が株式市場に反映されていくプロセスの解明に有用な情報となるものである。佐藤･高田(2017)で発表された。

(2) 長期高頻度株式統計の包括的解析

① 高頻度株式統計解析：
ニューヨーク証券取引所 TAQ 高頻度株式統計を用いて、重要変数についてノンパラメトリック適応的カーネル確率密度推計を行い、その形状変化パターンについて、トレンド転換予測可能性も含め、金融バブル前後に特徴的な、いくつかの事実を新しく発見した。Takada (2014)、岩本・高田(2015)で発表された。

② 高頻度板値統計解析：
ニューヨーク証券取引所 Openbook 指値注文統計を用いて、金融バブル崩壊前後の投資家行動の挙動の変化を売手/買手別に観察し、トレンド転換予測や、金融バブル生成メカニズム解明につながる、いくつかの統計的パターンについての新事実を発見した。Takada and Kitajima(2013)、北島・高田(2016)で発表された。

③ 人工知能育成による価格トレンド反転予測とトレンド反転への重要影響因子解明：
トレンド方向予測を行う人工知能を育成し、バイ・アンド・ホールド戦略よりもずっと高い収益性を挙げることに成功した。次に、その人工知能の内部を可視化し、どの因子がどのような形でトレンド反転に影響を与えているかを解析した。Takada and Kitajima (2016)で発表された。

５．主な発表論文等

〔雑誌論文〕（計６件）

① 佐藤圭、高田輝子、株式掲示板のトピック分析、OCU-GSB Working Paper, 査読無, 201702.

② Teruko Takada, Takahiro Kitajima, Robust Forecasting of Long-term smoothed trend reversals in stock market, 査読無, OCU-GSB Working Paper, 201608.

③ Yasutomo Tsukioka, Junya Yanagi, Teruko Takada, Investor sentiment extracted from internet stock message boards and IPO puzzles, OCU-GSB Working Paper, 査読無, 201506.

④ 佐藤圭、井上暁光、高田輝子、株式掲示板情報と株式市場の変動との関連性についての研究、OCU-GSB Working Paper, 査読無, 2015202.

⑤ 岩本菜々、高田輝子、高頻度株式データのノンパラメトリック確率密度推定による市場不安定性分析、OCU-GSB Working Paper, 査読無, 2015201.

⑥ Teruko Takada, Mining local tail dependence structures based on pointwise mutual information, Data Mining and Knowledge Discovery, 査読有, 24(1), 2012, 78-102.
DOI: 10.1007/s10618-011-0220-3

〔学会発表〕(計 13 件)

① 北島孝博、高田輝子、ノンパラメトリック確率密度推計による NYSE Openbook 売手/買手別指値注文パターンの非対称性分析、大阪大学中之島ワークショップ(招待講演)、2016 年 12 月 2 日、大阪大学中之島センター(大阪府大阪市).

② Teruko Takada, Yasutomo Tsukioka, Broken Corporate bond spread and investor risk appetite, The 9th CSDA International Conference on Computational and Financial Econometrics 2015 (Invited talk), December 12, 2015, University of London, London, United Kingdom.

③ Teruko Takada, Nonparametric density estimation based methods for robust risk analysis of trend reversals, Waseda International Symposium, November 10, 2015, Waseda University, Tokyo.

④ Teruko Takada, Robust risk analysis of abrupt switches in financial markets, Quantitative finance seminar (Invited talk), August 18, 2015, University of Zurich, Zurich, Switzerland.

⑤ Teruko Takada, Robust early warning signals of abrupt switches in stock markets, The 8th CSDA International Conference on Computational and Financial Econometrics 2014 (Invited talk), December 7, 2014, University of Pisa, Pisa, Italy.

⑥ 高田輝子、大規模金融データ解析で群衆行動の解明と制御を目指す、JST シンポジウム：情報学による未来社会のデザイン－人間力・社会力を強化する情報技術(招待講演)、2014 年 12 月 5 日、東京大学福武ホール(東京都文京区).

⑦ Teruko Takada, Takahiro Kitajima, Broken symmetry of financial bubbles: Evidence from NYSE limit order book, The 7th CSDA International Conference on Computational and Financial Econometrics 2013 (Invited talk), December 14, 2013, University of London, London, United Kingdom.

⑧ Teruko Takada, Robust big data analysis of financial bubbles, Kyoto University Informatics Seminar (Invited talk), July 8, 2013, Kyoto University, Kyoto.

⑨ 高田輝子、金融バブルの大規模データ解析、2013 年 6 月 7 日、2013 年度人工知能学会全国大会(招待講演)、2013 年 6 月 7 日、富山市民プラザ(富山県富山市).

⑩ 高田輝子、金融バブルのような相転移現象の予測法開発、競創ダイナミクスの統合的理解 第 1 回研究会(招待講演)、2013 年 3 月 8 日、名古屋大学(愛知県名古屋市).

⑪ Teruko Takada, Akimistu Inoue, Multiple time scale volatility patterns before abrupt switching in financial markets, The 6th CSDA International Conference on Computational and Financial Econometrics 2012, December 3, 2012, Conference center Ciudadde Oviedo, Oviedo, Spain.

⑫ 杉山将、Nigel Collier、高田輝子、大規模金融データによる不安伝播の解明と暴

走抑制のデザイン、３領域合同シンポジウム：情報学による未来社会のデザイン第１回「大量データに基づく未来社会のデザイン」（招待講演）、2012 年 11 月８日、一橋大学一橋講堂(東京都千代田区).

⑬ 高田輝子、佐藤圭、ウェブ検索頻度による投資家不安度の分析、2012 年度統計関連学会連合大会、2012 年 9 月 11 日、北海道大学高等教育推進機構(北海道札幌市).

６．研究組織

(1)研究代表者

高田 輝子（TAKADA, Teruko）

大阪市立大学・大学院経営学研究科・准教授

研究者番号：30347504