

ビットコイン市場におけるニュースの関係性における分析

Analysis on the relationship of news in bitcoin market

川田真也^{1,2,3} 尹熙元^{2,3} 藤原義久¹

Shinya Kawata^{1,2,3}, Hiwon Yoon^{2,3}, and Yoshi Fujiwara¹

¹ 兵庫県立大学大学院 シミュレーション学研究科

¹ University of Hyogo graduate school of simulation study

² 株式会社シーエムディーラボ

² CMDlab Inc.

³ 株式会社ビットアルゴ取引所東京

³ bitARG Exchange Tokyo Inc.

Abstract: In this research, we use English-language news related to the Bitcoin posted on the Internet as a data source, and use LDA(Latent Dirichlet Allocation) which is one of probabilistic topic models, for each article. We judged what kind of topics (keyword group) the sentence is composed of, and quantitatively expressed the excitement of topics in the period using different topic distribution for each article obtained. Furthermore, by analyzing the relationship with the bitcoin's market price on the Internet, we try to evaluate the influence of the news. We show to the relation between the quantity concerning the bitcoin's price (BTC/JPY) in the target period and the excitement of the topic in the news article related to bitcoin.

1. はじめに

日本では2017年4月1日に改正資金決済法が施行され、世界で初めて交換業としての仮想通貨の取り扱いは、国の認可が必要となった。そのことによって、世界から大きな注目を集めている。また法整備等に関しては、金融商品取引法ではなく、資金決済法の中で扱われている。その一方で、現状の仮想通貨は、従来の金融商品のような性質を持ち、投資や投機目的で用いられている。そのため、従来の金融商品であれば、禁止されているような相場操縦行為や風説の流布といった事象に抵触するような情報等が流れていることも目にする。現在の仮想通貨市場において、ある一つの話題やニュースなどが引き金となり、価格変動を助長しているようにも見受けられる。また情報のリソースがSNS等によるものであることを鑑みると、更新される新しい情報に対しての変化を捉え、健全な市場形成のために客観的な方法によって、提供される情報に関するガイドライン等の整備が必要であると考えられる。まず本研究では、仮想通貨に関する英字ニュースに対して、確率的トピックモデルの一つであるLDA(Latent Dirichle Allocation)を用いて、記事データに話題

の集積を数値的に評価し、価格データとの関係性について示す。

2. データについて

2.1. テキストデータ

今回は、インターネット上で公開されている英字の仮想通貨関連の記事を取得した。記事の取得期間は、2017年5月6日から2017年11月25日までとなる。さらに期間ごとに連続する(本研究では、2連続または、3連続)単語について、データセットごとに、TFIDFを用いて連続する単語の重要度を計算し、重要度の高い連続する単語を複合語として判断し処理を行う。

【複合語の例】

mining farms => mining-farms

bitcoin cash => bitcoin-cash

quantum computing => quantum-computing

initial coin offerings => initial-coin-offerings

・データ取得先

<https://blockchain.info/ja/charts/marketcap?timespan=all>

総記事数：1349 記事

総単語数：850836 単語 (ユニークな単語：60295 語)

2.2. ビットコイン[1]の価格データ

今回ビットコインの価格データに関しては、一日ごとの始値・終値・高値・安値を集計した国内の主要取引所の価格データを用いる。

3. 分析手法

3.1. 記事データにおける集積度

記事を用いた集積度（話題の集中度合）を定義する。単純な単語の出現数で判断することでは、同一のキーワードが出てきていないのもであっても内容については、同義の内容を示しているものなどの判定が難しくなる。そのため、確率的トピックモデルの一つのである LDA[2,3]によって、データセットごと（今回は、過去6週間の記事データを1データセットとした）に解析を行った。LDA に用いたパラメータは、今回データセットごとでの記事数の違いから一律のパラメータ（表1）を用いている。ただしパラメータの設定については、検討の余地を残している。LDA によって得られた記事ごとの話題の分布 $P(x)$ または $Q(x)$ から分布の類似度計算を行い、記事の話題の分布が近いものを特定することとした。記事間のつながり[4]の判定並びに類似度計算については、Jensen-shannon divergence を用いて計算を行った。以下に計算式(1)~(3)を表す。

<表1>LDA に用いたパラメータ

α (記事毎の話題の分布を推定に用いる)	0.25
β (話題内の単語の分布の推定に用いる)	0.01

$$D_{KL} = \int_x P(x)(\log P(x) - \log Q(x)) dx \quad (1)$$

$$M = \frac{P(x)+Q(x)}{2} \quad (2)$$

$$D_{js} = \frac{1}{2} D_{kl}(P||M) + \frac{1}{2} D_{kl}(Q||M) \quad (3)$$

得られた類似度計算から、データセット内でのすべての記事の組み合わせから類似度の高い組み合わせの上位1%を抽出する。抽出した組み合わせを用いてデータセット間の記事のつながりから記事のつながりでコンポーネントを判定し、指標（表2）を算出し、記事を用いた期間内での集積度の計算を行う。ただし、本研究においてコンポーネントは、グラフ構造内での分離されているネットワークを指すものとする。

<表2>集積度計算のための指標

$n_{i,a}$	データセット i における類似度計算で上位 1.0%に含まれる記事の総数 (コンポーネントが 2 つの記事で構成されているものは除く)
$n_{i,b}$	データセット i における記事数
C_i	データセット i のコンポーネントの数
R_i	データセット i の話題の集積度

$$R_i = \frac{n_{i,b}}{n_{i,a}} \times \frac{1}{C_i} \quad (4)$$

データセット内の記事数が 30 あるとし、記事の組み合わせから類似度の高い組み合わせの上位 1% を抽出した組み合わせから作成した記事ネットワーク（図1）について考える。

article00~article05,article07 は一つのコンポーネント、さらに article_11~article13 も一つのコンポーネントを形成している。定義より article09, article10 (article14, article15) はコンポーネントであるが二つの記事のみでのコンポーネントであるために除外するとする。

この結果、各指標は表3のようになる。

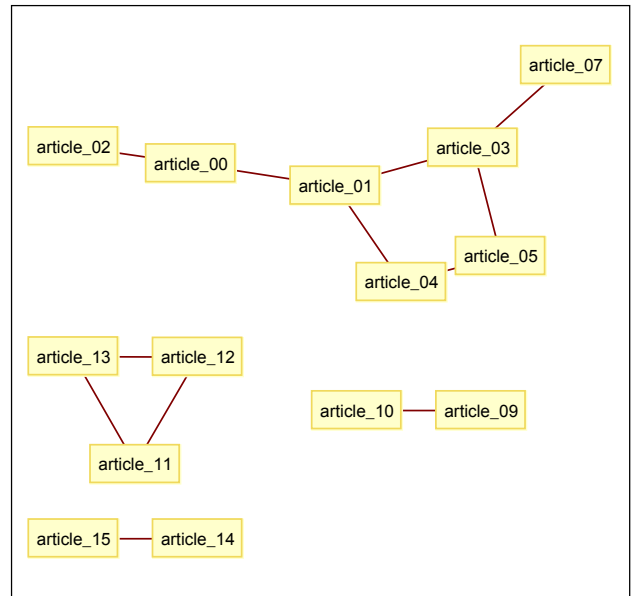


図1：記事ネットワーク

<表2>集積度計算のための指標

$n_{i,a}$	10
$n_{i,b}$	30
C_i	2
R_i	1.50

3.2. ビットコインの価格データの処理について

今回ビットコインの価格データに関しては、一日ごとの始値・終値・高値・安値を集計したデータから1週間ごとの始値・終値・高値・安値に再集計したデータを用いる。

4. 結果及び考察

ビットコインの価格(図2)が2017年1月から2017年11月にかけて5倍近く、その他の仮想通貨でも軒並み同様かそれ以上の価格の上昇が見られることを鑑みると価格の推移を比較するのではなく、1週間ごとの高値・安値から1週間ごとの高値・安値を計算し、価格変動幅を求め停止した話題の集積度との比較を行った。

価格変動幅の変化率と話題の集積度の変化率との相関を取ると0.49071となり弱い相関を持つ。同様の解析を対象とするデータセットの期間を1週間から2ヶ月周期で行なった際には、相関が、0.813と高い相関を持つことはわかっている。

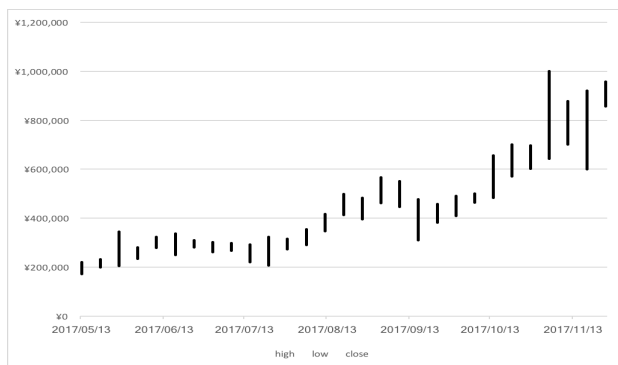


図2：ビットコインの価格(BTC/JPY)の週足データ (2017/5/6-2017/11/24)

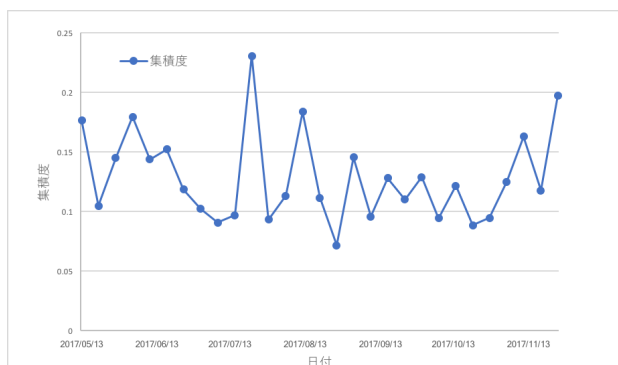


図3：話題の集積度の週ごとの推移 (2017/5/6-2017/11/24)

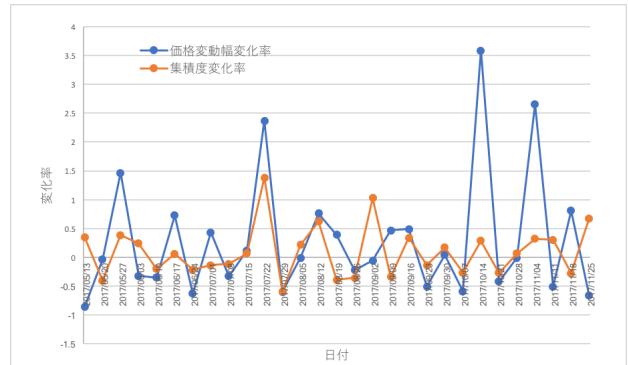


図5：価格変動幅と集積度の変化率

単純な相関のみを評価したが、実際にテキスト情報と価格変動の因果関係については、さらなる検討が必要になる。因果関係を考えるためには、正確にどのタイミングで情報を受け取っているのかという点も考慮する必要がある。

また、ビットコインの価格のように変動が激しい時系列データに関しての処理等についても考慮する必要がある。

5. 今後の展望

本研究では、英字テキストとビットコイン価格(BTC/JPY)を用いての解析であったが、日本語によるデータソース及び国内の主要取引所等でのビットコイン価格(BTC/JPY)に関しての解析を今後行うこととしている。また単語数がある程度見込めるリソースでの解析であるが、実際の市場を鑑みると短文形式のSNSや仮想通貨に関するまとめサイトさらには、市場への注目度を鑑みると一部の専門的なメディアだけではなく、様々な媒体に対してのアプローチ方法について現在検討中である。

謝辞

本研究は、株式会社シーエムディーラボ並びに株式会社ビットアルゴ取引所東京(仮想通貨交換業 第00011号 関東財務局)によるデータの提供、協力によって行うことができました。ここに深く感謝いたします。

参考文献

- [1] Nakamoto, S. Bitcoin: A peer-to-peer electronic cash system. (2008).
- [2] Blei, D. M., and Ng, A. Y. and Jordan, M. I., Latent Dirichlet Allocation, Journal of Machine Learning, 3, 993-1022 (2003)
- [3] Phan, X.-H. and Nguyen, C. T., GibbsLDA++: A C/C++

Implementation of Latent Dirichlet Allocation(2008)

- [4] Kawata, S., & Fujiwara, Y.. Constructing of network from topics and their temporal change in the Nikkei newspaper articles. *Evolutionary and Institutional Economics Review*, 13(2), 423-436(2016).