Twitter のテキストとネットワークの解析による経済動向分析

Economic Trend Analysis by Text-Mining and Network Analysis of Twitter

迫村 光秋[1]　和泉 潔[1,2]　セーヨー サンティ[3]

Mitsuaki Sakomura[1], Kiyoshi Izumi[1,2], Santi Saeyor[3]

[1] 東京大学大学院 工学系研究科
[1] School of Engineering, The University of Tokyo
[2] 科学技術振興機構 さきがけ&CREST
[2] PRESTO & CREST, JST
[3] 株式会社ホットリンク※
[3] Hottolink, Inc.

1. 緒言

1.1 背景

現在ツイッターなどのマイクロブログには、様々なニュースとそれに対する人々の反応が書かれており、その情報量は莫大かつ増加し続けている[1]。

この莫大な情報を実世界の動きを観測するためのソーシャルセンサとして利用する研究の数は増加しており、観測する対象をあらかじめ設定し、それについて詳細な分析を行ったものが多くみられる[2]。中でも、経済動向を分析対象としたものとして、ツイッターからキーワードを用いて特定の株式銘柄に関する情報を収集し、株価動向との関連分析に取り組んだ事例[3]があるなど、ツイッターは経済動向の分析に大いに用いられている。

1.2 既存研究

Bollen ら[4]は、ツイートの内容を対象に気分プロフィール調査を行うことで、「平穏」、「警戒」などの 6 つの心的状態を表す指数を抽出し、ダウ平均株価の予測を行った。しかし、分析対象となるツイートは“I feel”, “I’m”といった心的状態を明言したものに限られていることに加えて、ツイート情報はダウ平均株価の過去の数値データによる予測を補うものとして用いられている。

Ruiz ら[5]は、ツイッター情報から特定のキーワードを用いて株価を予測する銘柄に関連するツイートを抽出し、ツイート数やユーザー数などの活動基準の特徴量とリツイートやユーザーへの言及などをグラフ表現した際のノード数やエッジ数といったグラフ構造特徴量の 2 つの特徴量と株価、出来高との相関を調べた。しかし、ツイート本文の情報は分析されていない。

1.3 本研究の目的

本研究ではツイッター情報からテキストの特徴量とグラフ特徴量の 2 つの特徴量を抽出し、得られた情報と経済動向との関連性を明確にすることで、ツイッター上の膨大な情報の中から経済動向の分析に有用な情報を得ることを目的とする。

テキストの特徴量とは、ツイート内容に含まれる単語の頻度を基準に算出さるものであり、ツイートの話題を示す。グラフ特徴量とは、ツイッターをグラフ表現したときの特徴量であり、ツイートの話題の大きさや広がり方を示す。

この 2 つの特徴量を用いることで、ツイッター上で話題となっている内容とその話題の広がり方の 2 つと、経済動向との関連性を明らかにする。

具体的には、ツイッターから抽出された特徴量により経済指標の予測を行う。そして、各特徴量について経済指標の予測に与えた影響の大きさを比較することで、その影響度の違いを明らかにする。

2. 分析手法

2.1 分析手法の全体像

図 1 に本研究で新たに開発した分析手法の全体像を示す。分析手法の概略を以下に述べる。

1. ツイッターから経済に関連するツイートを抽

※ 本研究の内容は株式会社ホットリンクの公式見解を示すものではありません。

出する。
2. 1.のツイートに対してテキスト分析を行い、テキストの特徴量を抽出する。
3. 1.のツイートに対してグラフ表現分析を行い、グラフ特徴量を抽出する。
4. 2. 3.により抽出された特徴量と経済動向を表す指数から回帰分析を行う。
5. 外挿予測を行い、経済動向の分析を行う。

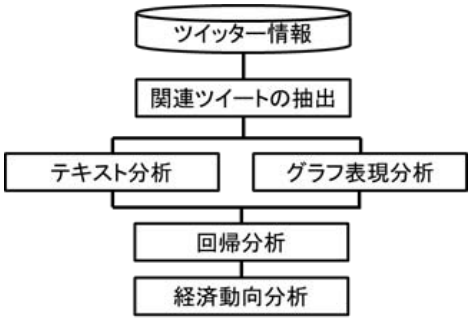

図 1　分析手法の全体像

なお、本研究ではツイートは日次でまとめて分析した。

3. テキスト分析

3.1 関連ツイートの抽出

まず、ツイート内容に対して形態素解析を行い、日経シソーラス[2]に収録されている単語のうち、経済と特に関連が高い分野の単語を含むツイートを抽出し、分析対象とした。

具体的には、日経シソーラスの単語分野一覧から、「経済動向、見通し」、「金融一般」などの 13 分野に含まれる単語を含むツイートを経済に関連するツイートとして抽出した。

形態素解析には高速全文検索エンジン Lucene 用の日本語形態素解析プラグイン lucene-gosen[3]を利用した。

なお、形態素解析の辞書には、日経シソーラスに含まれる単語を追加した。日経シソーラスは日本経済新聞デジタルメディアが作成している新聞記事検索のための用語集であり、14,879 語が収録されている。

それに加えて、ツイッターというインターネット上の口語体の文章に対応するために、はてなキーワード[4]から作成した辞書も合わせて追加した。はてなキーワードは株式会社はてなが提供する共有辞書サービスであり、流行語やインターネット上での俗語表現、固有名詞、複合語に強い。本研究において、はてなキーワードから作成した辞書の登録単語数は 244,076 語である。

3.2 単語出現頻度行列の作成

次に、分析対象のツイートに含まれる名詞・動詞・形容詞の単語出現頻度を数え上げ、これを日次のツイートで繰り返し行った。これにより得られた単語出現頻度を、時系列順に、後述する回帰分析の訓練期間である 30 日分並べた単語出現頻度行列を作成した。

なお、ツイッター全体の成長による影響を除くために、式(1)により、ある日付における単語の出現頻度を、その日付において分析対象としたツイートの数で割った値を用いた。

$$WordFreq_{i,t} = \frac{WordCount_{i,t}}{NumOfTweet_t} \qquad (1)$$

$WordCount_{i,t}$　：日付tにおける単語iの出現回数

$NumOfTweet_t$　：日付tにおける分析対象としたツイート数

この単語出現頻度の最小値や単語が出現した日数に閾値を設けることで、ほとんど出現しない単語を除外した。本研究では、単語出現頻度の最小値を 0.0001 とし、単語出現頻度行列を作成する期間の半分以上の日数に出現した単語を対象とした。

3.3 主成分分析

3.2 で得られた 30 日間の単語出現頻度行列に対して、主成分分析を行った。

新聞記事やオンラインニュースといった定型的な文章とは異なり 140 文字という限られた分量からなるツイートでは様々な表現があるが、主成分分析を行うことにより単語をグルーピングすることができる。この利点は、各単語に対してパターンマッチングを行い、頻度を算出する従来の手法では取りこぼしてしまう情報を、主成分としてまとめることで有効に活用できることである。

それに加えて、各主成分において因子負荷量の絶対値が大きい単語はその主成分を表す主要単語として考えることができる。これにより、人が主要単語の一覧を把握することで、それぞれの主成分に話題やトピックなどの意味を与えて解釈することができる。

主成分の数は累積寄与率が 80%を超える 15 個とし、各日の主成分スコアを算出することで、15 次元からなるベクトルにより 1 日のツイートを評価するテキストの特徴量を得た。

[2] http://t21.nikkei.co.jp/public/help/contract/price/01/help_kiji_thes_field.html
[3] http://code.google.com/p/lucene-gosen/
[4] http://d.hatena.ne.jp/hatenadiary/20060922/1158908401

4. グラフ表現分析

4.1 グラフのモデル

3.1 によって抽出された経済に関連するツイートを対象にグラフ表現分析を行う。図 2 にグラフのモデルを示す。

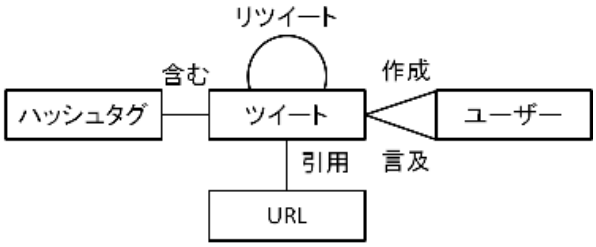

図 2　ツイートのグラフモデル
(Ruiz らの研究[5]より作成)

このグラフのモデルは Ruiz らの研究で提唱されたものであるが、分析対象とするツイートの抽出方法にハッシュタグやティッカーコードを用いるのではなく、3.1 で述べたようにツイートの内容によって抽出しており、データの規模は大きく異なる。

表 1 にこのモデルによって表現されたグラフのノードとエッジの一覧を示す。

表 1　グラフのノードとエッジの一覧

ノード	説明
ツイート	1つのツイートを示す
ユーザー	ツイートしたユーザー/ツイートに含まれるユーザー
URL	ツイートに含まれるURL
ハッシュタグ	ツイートに含まれるハッシュタグ

エッジ	説明
含む	ツイートがハッシュタグを含む
リツイート	リツイートの関係
作成	ユーザーがツイートを作成する
言及	ツイートでユーザーについて言及する
引用	ツイートにURLを引用する

このモデルを用いて、3.1 で抽出した日次のツイートをグラフ表現した。図 3 にツイートをグラフ表現したネットワーク図の例を示す。

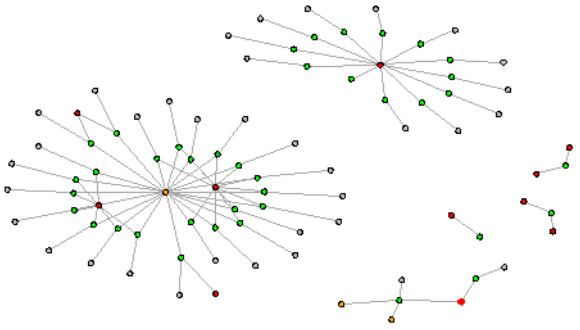

図 3 ツイートをグラフ表現したネットワーク図

図 3 は、2011 年 9 月 1 日のツイートの一部を用いて作成したものである。ノードは丸、エッジは線で表されており、各ノードについて、ツイートを緑、ユーザーを赤、ハッシュタグを黄、URL を灰色の色別で表現している。

4.2 グラフ特徴量の算出

このように日次のツイートをグラフ表現した後、活動基準の特徴量とグラフ構造特徴量の 2 つからなるグラフ特徴量を算出した。表 2 にそれぞれの特徴量の一覧を示す。

表 2　グラフ特徴量の一覧

活動基準の特徴量	説明
ツイート数	ツイートの数を示す
ユーザー数	ユニークユーザー数を示す
URL数	ユニークURL数を示す
ハッシュタグ数	ユニークハッシュタグ数を示す

グラフ構造特徴量	説明
ノード数	ノードの総数を示す
エッジ数	エッジ数を示す
平均次数	1ノードあたりの平均次数を示す
連結成分数	連結成分の数を示す

本研究では、経済情報を含むツイートを抽出しているため、株式市場が開いている平日と、そうでない休祝日とでは、グラフ表現した際のエッジ数とノード数の両方に大きな違いがあり、カレンダー効果が確認できた。この影響を取り除くため、休祝日のツイートを分析対象から除外した。

また、回帰分析を行う際に説明変数として用いるため、それぞれの説明変数のオーダーが大きく異ならないように、活動基準の特徴量とグラフ構造特徴量を平均が 0、分散が 1 となるように標準化した。

5. 回帰分析

5.1 経済指標

本研究では、経済指標として TOPIX、日経平均株価、業種別日経平均株価の 3 つの株価指数を利用した。

なお、回帰分析の被説明変数として、これら各経済指標の日次の変動率を示す絶対リターンを用いた。経済指標i の日付t における終値を$p_{i,t}$とすると絶対リターン$r_{i,t}$は式(2)で定義できる。

$$r_{i,t} = \frac{p_{i,t+1} - p_{i,t}}{p_{i,t}} \qquad (2)$$

5.2 回帰式の作成

回帰分析の説明変数として、3.3 により得たテキストの特徴量と 4.2 により得たグラフ特徴量の時系列データを用い、被説明変数として、5.1 で示した各経済指標の絶対リターンを用いることで、線形回帰式(3)を作成した。

$$r_{i,t} = a_{i,0} + \sum_{j=1}^{n_{pc}} a_{i,j} x_{j,t} + \sum_{k=1}^{8} b_{i,k} y_{k,t} \quad (3)$$

$r_{i,t}$ ：時刻tにおける指標iの絶対リターン
$a_{i,j}$ ：指標iの回帰式におけるテキストの特徴量jの回帰係数
$x_{j,t}$ ：時刻tにおけるテキストの特徴量j (主成分jのスコア)
$b_{i,k}$ ：指標iの回帰式におけるグラフ特徴量kの回帰係数
$y_{k,t}$ ：時刻tにおけるグラフ特徴量k

回帰式の訓練期間は 30 日間とした。また、回帰式を作成する訓練期間の開始日を 1 日ずつ移動していき、その度にテキスト分析とグラフ表現分析を行うことで回帰式を更新した。

なお、回帰分析では AIC 基準のステップワイズ変数選択を行い、説明力の低い変数は回帰式から除外した。

また、回帰分析によって得られた各経済指標の予測変動率について、その変動の方向性が一致した場合は正解、不一致の場合は不正解として予測正答率を算出した。外挿期間は、利用したツイッターデータの期間から訓練期間と祝祭日を除いた 30 日間である。

6. 予測実験

6.1 実験データ

株式会社ホットリンク[5]より提供されたデータを利用した。データは、日本人であろうアカウント約 900 万を特定し、そのアカウントごとのツイートを巡回的に収集して取得したものである。データの概要を表 3 に示す。

表 3 データの概要

期間	2011/9/1～2011/12/31
ツイート数	約32億件
ユーザー数	約900万
データ容量	約2.2TB

6.2 実験環境

本研究で扱うツイッター情報は 2TB 以上の大規模なデータであり、分散処理を行う必要がある。そこで、Google が開発した大規模データを並列分散処理するためのアルゴリズム MapReduce[6]をもとに、開発された Hadoop[7]を利用した。Hadoop とは大規模データを効率的に分散処理するための Java ソフトウェアフレームワークである。

本研究で使用した Hadoop クラスタの概要としてクラスタの各マシンの役割を表 4 に示す。

表 4 Hadoop クラスタの各マシンの役割

マシンNo.	役割	説明
1	Name Node, Data Node	データの分散管理、データの保持、分散処理
2,3,4	Data Node	データの保持、分散処理

6.3 テキスト分析による予測

ツイッターデータに対して、テキスト分析だけを行い、テキストの特徴量を抽出し、回帰分析によって経済動向を表す指標の予測を行った。

表 5 に 2011 年 10 月 12 日から 30 日間のツイートをテキスト分析して得た主成分の例を示す。

表 5 主成分を表す主要単語の例

主成分	主要単語（因子負荷量の降順）				
PC1(+)	わい	ランキング	輝	書評	マニュアル
	位	喉	確定申告	ムック	ラップ
PC1(−)	よう	震災復興	そう	お互い	助ける
	容認	暴動	韓国	援助	血税
PC2(+)	在庫	ビタミン	検察	会計監査	案内
	製品	医師	久しぶり	ムービー	者
PC2(−)	殴る	最上級	しか	位	乏しい
	ふる	X777	日本郵政	RI	さき
PC3(+)	全国	TPP	米国	反対	対象
	市場原理	外	参加	圏外	支払
PC3(−)	ウォレット	愛顧	サマリー	楽天	発売
	株式投資	静岡県	花の慶次	ガーデン	ニュース速報
PC4(+)	投機	円高	一時	円相場	安価
	遭う	急激	入金	東京市場	最高値
PC4(−)	軽減	機能性	情報局	低金利	党
	攻略法	融資	キャッシング	甘い	金融情報

なお、各主成分について因子負荷量が正負それぞれ絶対値が大きい上位 10 単語を主要単語と定義した。

PC3(+)を見ると、「TPP」、「米国」、「反対」といった TPP に関する話題を示すことがわかる。PC4(+)を見ると、「投機」、「円高」、「円相場」など為替相場に関する話題を示すことがわかる。このように、ツイッターデータから経済に関連する話題を抽出できたことが確認できる。

6.4 グラフ表現分析による予測

次にグラフ表現分析を行い、グラフ特徴量を抽出し、回帰分析によって経済動向を表す指標を予測した。本研究で分析対象とした経済に関連するツイートをグラフ表現した場合の、次数の大きさ上位 15 ノードの一覧を表 6 に示す。なお、表 6 は 2011 年 11 月 19 日のツイッターデータから作成したものである。

表 6 次数の大きさ上位 15 ノードの一覧

1	#followmeJP	6	@rakuten_realtim	11	#niconews
2	#sougofollow	7	#followme	12	#TPP
3	#2ch	8	#bot	13	#seiji
4	#followmejp	9	#ogiri_111119	14	#Amazon
5	URL:http://asahi.com	10	@ogiri_tweet	15	@Yomiuri_Online

#seiji、#TPP、@Yomiuri_Online などは政治に関するハッシュタグ、ニュースサイトのユーザーを示しており、経済に関連するツイートのグラフ構造が抽出されていることが確認できる。

[5] http://www.hottolink.co.jp/
[6] http://research.google.com/archive/mapreduce.html
[7] http://hadoop.apache.org/

6.5 テキスト分析とグラフ表現分析による予測

テキスト分析とグラフ表現分析の 2 つの分析を行い、テキストの特徴量とグラフ特徴量を抽出し、これらの 2 つの特徴量を用いて回帰分析を行うことで経済動向を表す指標を予測した。

6.6 予測結果と考察

図 4 に、分析手法による市場平均指標の予測正答率の違いを示す。

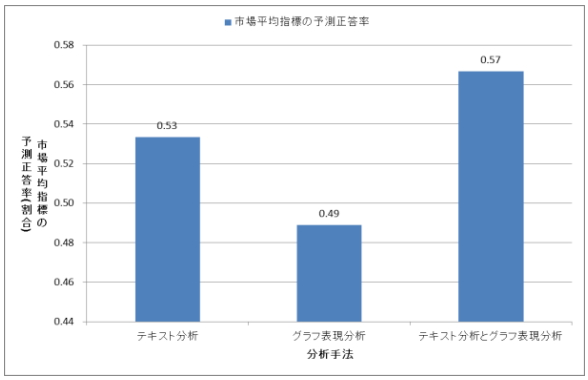

図 4　分析手法による市場平均指標の予測正答率の違い

経済全体の動向を表す市場平均指標である日経平均と TOPIX については本研究で開発したテキスト分析とグラフ表現分析による予測の正答率が最も高かった。また、回帰式の当てはまりを示す自由度調整済み決定係数についても、同様に最も高くなった。

図 5 に、分析手法による業種別指標の予測正答率の違いを示す。

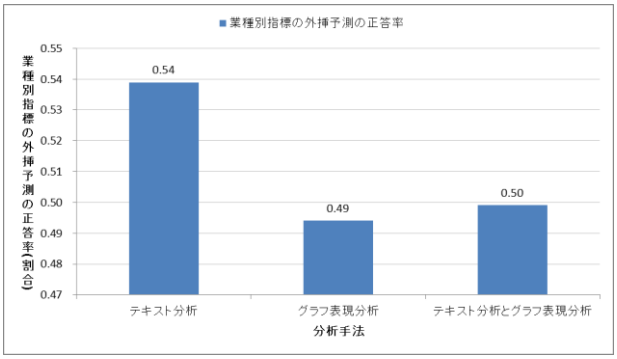

図 5　分析手法による業種別指標の予測正答率の違い

テキスト分析による業種別指標の平均予測正答率は 0.538 であり、二項検定を行ったところ、ランダムで予測した場合の正答率 0.5 と有意水準 1%で差が見られた。

また、市場平均指標とは異なり、業種別指標の平均予測正答率はテキスト分析とグラフ表現分析を組み合わせても、正答率は上昇しなかった。

この理由として、テキストの特徴量が経済に関連する話題を示し、グラフ特徴量がその話題の広がり方や大きさを示しているため、経済全体を表す市場平均指標の予測ではグラフ特徴量が有用であったが、業種別の予測では、経済全体の影響を与えることとなり、正答率が低くなったと考えられる。一方、テキストの特徴量は、主成分ごとに TPP や為替など経済に関する具体的な話題を示しており、経済全般の予測だけでなく、業種別の予測にも有用であった。

また、内挿予測の平均正答率は 0.860 と高く、過去の訓練期間におけるテキストの特徴量やグラフ特徴量の回帰係数を比較することにより、どのような特徴量がどの程度予測に影響を与えていたのかを分析することができる。

7. 結言

本研究では、ツイッターデータからテキストの特徴量とグラフ特徴量の 2 つの特徴量を抽出し、経済動向を表す指標との関連性を明らかにするテキスト分析、グラフ表現分析、回帰分析からなる分析手法を開発した。

また、実際のツイッターデータに対して、本研究で開発した分析手法を用いることで、経済指標の予測を行った。そして、テキスト分析とグラフ表現分析、両者を組み合わせた分析のそれぞれの予測結果を比較することで、予測する経済指標によるテキストの特徴量とグラフ特徴量の影響の違いを明らかにした。

参考文献

[1] Twitter Inc., “Twitter Blog: Twitter turns six”,2012 年 3 月 21, htttp://blog.twitter.com/2012/03/twitter-turns-six.html (2012 年 8 月参照)

[2] 榊剛, 松尾豊, “ソーシャルセンサとしての Twitter: ソーシャルセンサは物理センサを凌駕するか?”, 人工知能学会誌, 27 巻, 1 号, pp.67-74,2012

[3] 日本アイ・ビー・エム株式会社, “株ドットコム証券株式会社 谷口有近氏 インターネット上の膨大なデータを収集・分析し、株価との関連性に基づいた新サービスの提供を模索”, PROVISION72 号, pp16-23, 2012 http://www-06.ibm.com/ibm/jp/provision/no72/pdf/72_interview1.pdf (2012 年 12 月参照)

[4] Bollen, J., Mao, H. and Zeg, X. “Twitter mood predicts the stock market.” J.computational Science, Vol.2, No.1,

pp.1-8, 2011

[５] Eduardo J. Ruiz, Vagelis Hristidis, Carlos Castillo, Aristides Gionis, and Alejandro Jaimes. “Correlating financial time series with micro-blogging activity.”, ASDM, 2012