twitter テキストマイニングによる経済動向分析

Economic trend analysis by text-mining of twitter

迫村光秋 [1]　和泉潔 [1]

Mitsuaki Sakomura[1], Kiyoshi Izumi[1]

[1] 東京大学
[1] The University of Tokyo

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、様々なニュースやそれに対する人々の反応が書かれており、その情報量は膨大かつ、増加し続けている[1]。この膨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ーとして利用する研究の数は増加しており、観測する対象を予め設定し、それについて詳細な分析を行ったものが多く見られる[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 46 の特定の株式銘柄に関する情報を収集し、株価動向との関連の分析に取り組んだ事例[3]があるなど、ツイッター情報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

既存研究として、Bollen ら[4]は、ツイートを対象に気分プロフィール調査を行い、「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、ツイー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情報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情報量の 2 つ情報量と株価、出来高との相関を調べた。しかし、ツイートに含まれる単語などの情報は分析されていない。

このように、ツイッター情報から経済動向の分析を行った研究はあるものの、Bollen ら[4]の研究では、心的状態を明言したものだけに対象を限定したうえで、ツイートの内容のみを分析しておりツイッターのグラフ構造については触れられていない。それに対して、Ruiz ら[5]の研究では、ツイッターのグラフ構造について分析しているものの、ツイート内容については触れられていない。

1.3 本研究の目的

本研究では、ツイッター情報から短期的な経済動向の分析を行うことを目的とする。ツイッターからツイート内容に含まれる単語をベースとしたテキストの特徴量とグラフ表現した際のグラフ特徴量の 2 つを抽出することで、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益な情報を得る。

テキストの特徴量を調べることで、人々がツイートする内容と経済動向との関連性を明らかにし、グラフ特徴量を調べることでツイートの広がり、すなわち、話題性の大きさを抽出できる。これらを組み合わせることで、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、TOPIX、日経平均株価、日経 500 種平均株価（業種別日経平均株価）を経済動向の分析対象とし、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、投資活動に役立つ情報を提供する。

2. データマイニング手法

データマイニングの手法は大きく次の 2 つに分類される。ツイッターからテキストの特徴量を抽出するテキスト分析とグラフ特徴量を抽出するグラフ表現分析である。これら 2 つの特徴量を抽出した上で、1.3 で述べた経済動向を表す指数との回帰分析を行う。図 1 にデータマイニング手法のフローを示す。

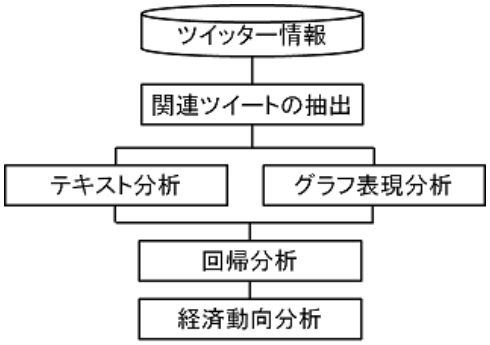

図 1　データマイニング手法のフロー

ツイートは日次でまとめて分析する。

2.1 関連ツイートの抽出

まず、ツイートに対して形態素解析を行い、動詞・名詞・形容詞を抽出する。そして日経シソーラス[6]に収録されている経済専門用語を含むツイートのみを分析の対象とする。

次に、分析対象のツイートに含まれる単語の出現数を数え上げ、これを日次のツイートで繰り返し行い、訓練期間日数分の単語の出現パターン行列を作成した。なお、ツイッター全体の成長による影響を除去するために単語の出現数を対象日の全ツイート数で除した値を用いた。

形態素解析には高速全文検索エンジン Lucene[7]用の日本語形態素解析プラグイン lucene-gosen[8]を用いて、日経シソーラスに含まれる単語を形態素解析の辞書に追加した。

なお、日経シソーラスと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、約 1 万 3 千語が分野別に収録されている。TOPIX、日経平均株価を予測する際には日経シソーラスの全単語を利用し、業種別日経平均株価を予測する際には、予測対象となる業種に関連する分野の単語のみを利用した。

このように経済専門用語を含むツイートに対象を限定することで、経済動向とは関連のないツイッター上の多くのノイズを除去することができる。

2.2 主成分分析

日次で数え上げた訓練期間日数分の単語の出現パターン行列に対して主成分分析を行い、30 の主成分でツイート内容を評価した。

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ため、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。すなわち、経済専門用語を辞書としたパターンマッチングでは見落とされる単語を、関連する経済専門用語と同一の主成分に含ませることができる。

2.3 グラフ表現分析

2.1 によって抽出した関連ツイートを対象にグラフ表現を行う。図 2 にグラフモデルを示す。

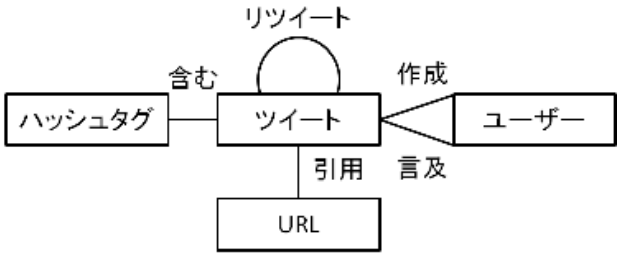

図 2　ツイートのグラフモデル

このモデルは、Ruiz ら[5]の研究で用いられたものであるが、対象とするツイートを単純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、2.1 で述べたようにツイートの内容によって抽出しており、範囲は大きく異なるものである。

表 1 に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

このモデルによって、2.1 で抽出した日次のツイートをグラフ表現し、活動基準の特徴量とグラフ特徴量の 2 つの特徴量を算出する。表 2 にそれぞれの特徴量の一覧を示す。なおこれらの特徴量もツイッター全体の成長による影響を除去するために、正規化した。

表 2　グラフの特徴量の一覧

行動基準の特徴量
ツイート数
ツイートしたユーザー数
リツイート数
リツイートしたユーザー数
ユーザーに言及したツイート数
URLを引用したツイート数

グラフ特徴量
ノード数
エッジ数
コンポーネント数

2.4 回帰分析

2.2 にて算出した主成分スコアと 2.3 にて算出したグラフ特徴量を説明変数とし、1.3 にて述べた株価指数の変動率を被説明変数とした。

株価指数 i の時刻 t における値を p_(i,t)とすると変動率 r_(i,t)は下式で定義できる。

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

訓練期間中の各日の主成分スコア x_(j,t)とグラフ特徴量 y_(j,t)を用いて下式に示す重回帰分析を行う。

$$r_{i,t} = a_{i,0} + \sum_{j=1}^{20} a_{i,j} x_{j,t} + \sum_{j=21}^{29} a_{i,j} y_{j,t} \qquad (2)$$

回帰分析の訓練期間は 30 日間で、直近の単位期間後の予測を行う。単位期間⊿t は 1 日である。

回帰式を推定する際には AIC 基準においてステップワイズ選択を行うことで、説明力の低い変数を回帰式に含めず、モデルの複雑性を制御している。

3. 予測実験

3.1 実験環境

本研究で扱うツイッター情報は数百 GB 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[9]をもとに、オープンソースで実装された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 Hadoop[10]を利用した。

表 3 に本研究で利用した Hadoop クラスタのマシンのスペックを示す。本研究で用いたクラスタは、分散管理用のマスターノード 1 台と分散処理用のスレーブノード 3 台の計 4 台で構成した。

表 3　マシンのスペック

CPU	Intel(R) Core(TM) i5-2500K CPU @ 3.30GHz
Memory	16GB
HDD	2TB*2

3.2 実験データ

データの概要を次に述べる。期間は 2011 年 9 月から 2011 年 12 月までの 4 ヶ月、アカウント数は 300 万、容量は約 20GB/日である。

具体的なデータの内容として、ツイート本文に加えて、screen_name、id、created_at といったユーザーID や作成日時を示す一般的 tweet ステータスとリツイート関係を示すリツイートステータスが含まれる。

参考文献

[1] Twitter ホームページ
http://blog.twitter.com/2012/03/twitter-turns-six.html

[2] 榊剛史、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] IBM ホームページ
http://www-06.ibm.com/software/jp/casestudies/pdf/20111226kabudotcom.pdf

[4] Bollen, J., Mao, H. and Zeng, 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.WSDM, 2012

[6] 日経テレコン 21 ホームページ　日経シソーラス
http://t21.nikkei.co.jp/public/help/contract/price/01/help_kiji_thes_field.html

[7] 高速全文検索エンジン Lucene ホームページ
http://lucene.apache.org/core/

[8] 日本語形態素解析プラグイン lucene-gosen ホームページ
http://code.google.com/p/lucene-gosen/

[9] J.Dean and S. Ghemawat MapReduce: Simplified Data Processing on Large Clusters. 2004
http://research.google.com/archive/mapreduce.html

[10] Hadoop ホームページ
http://hadoop.apache.org/