12-DA05

大規模テキストを利用した経済指標分析手法に関する研究

和泉　潔（東京大学大学院 工学系研究科）

概要
経済時系列データの分析の手法の一つとして、関連するテキストを用いて将来の時系列値を予測するという手法がある。本研究課題では、予測精度向上に向けて、「言語的資源の利用」「大量テキストに対する機械学習手法の改善」を目的とした様々な手法の検討を行った。具体的には、機械学習におけるパラメータ調整の効果の検証、次元削減手法の適用および重要単語抽出手法の開発、言語資源を利用した予測手法の開発を行った。

1. 研究の目的と意義

本研究では、経済時系列データの分析を大規模に行うことを目的とする。特に、テキストデータからの経済指標（株価や取引高）の予測、経済指標を利用した言葉の意味の獲得など、テキストと経済時系列データとの関連を分析することを主眼に置く。

従来、経済指標の分析は、経済指標そのものの時系列データ（cf.テクニカル分析）等、数値的なデータを元にした手法が主であったが、近年、電子的テキストの普及に伴い、テキストデータを経済指標の分析や予測に用いる手法についての研究が盛んになりつつある。特に、WWW には、ニュース、株式情報の掲示板、ブログ、ソーシャルメディア(Twitter 等)といった、様々な媒体でテキスト情報が提供されており、そのデータ量の爆発的増大が、計算機によるテキストデータの取り扱いを非常に重要なものとしている。具体的な利用例としては、テキストデータを自動的に解析することで、取引戦略の決定を行う、あるいは、人が戦略の決定を行う意志決定を支援する、といったものが挙げられる。例えば、テキストデータを分析することで、ある銘柄の株価が上昇する確率が高いか、下落する確率が高いかを判断できれば、取引戦略の決定に有用である。その他にも、取引高を予測することで、その日に流動性が高くなる銘柄を予測する等の応用も考えられる。

また、テキスト分析の観点からは、経済的テキストの分析や、経済指標を利用した意味抽出は、数値時系列データとテキストの関連分析として捉えることができる。これは、特に、多種多様な数値データ・テキストデータが WWW を通じてアクセスできる今日的な状況において近年注目されており、開拓が期待される分野である。

本提案では、以下の３つの観点から、この分野の研究を発展させることを目的とする。

(1)テキストデータに対する詳細な言語的分析を行うことによる、経済指標予測精度の向上

経済用語を収録した辞書(「日経シソーラス」等)、銘柄の業種別分類、言語解析ツール（係り受け解析・固有表現抽出器「Cabocha」等）の利用法・有用性について検討を行う。また、テキストマイニング技術を応用した特徴量抽出等の検討を行う。

(2)大量テキストの解析による予測精度の向上

・大量のテキストデータを利用することにより、データ量に応じて株価予測精度が向上するか否かの調査を行う。そのための、大量テキストを高速に処理し分析を行う手法を開発する。

(3)オンライン学習を利用したテキストの効率的利用方法の確立

新聞記事のような、時系列的に到着するデータを効率的に利用するためには、近年注目を集めるオンライン学習手法が有効である。このオンライン学習の、経済テキスト分析への応用についても研究を進める。

2. 当拠点公募型共同研究として実施した意義

(1) 共同研究を実施した大学名と研究体制

東京大学（代表者：和泉潔、副代表者：吉田稔、共同研究者：中川裕志）
中部大学（共同研究者：松井藤五郎）

(2) 共同研究分野

超大規模データ処理系応用分野

(3) 当公募型共同研究ならではという事項など

本研究は、テキストを高速に処理し、経済指標解析に活用することを目指しているため、テキストを前処理し、機械での数値処理に適した形に変換するための自然言語処理技術、また、それを高速に処理するためのテキストマイニング技術が必要となる。学術情報研究部門の、自然言語処理、および、テキストの高速処理における専門家の助言により、この処理を高度化できる。

3. 研究成果の詳細と当初計画の達成状況

(1) 研究成果の詳細について

(1)言語的資源を用いた予測精度向上に関する研究

新聞記事を対象として長期的な経済動向の分析を行うため、言語資源を利用した予測手法の開発を行った。

分析手法

本研究では CPR 法を用いて分析を行う．CPR 法とは共起解析，主成分分析，回帰分析の三段階からなる分析手法である[和泉 2011]．従来の CPR 法で入力テキストとしていた日本銀行の金融経済月報は比較的形式の決まった文書であるため，非常に扱いやすいものであった．本研究では新聞記事という形式が金融経済月報よりも定まっていない文章を用い，より広範で長期的な分析が可能となるように CPR 法を拡張した．

共起解析

本研究で用いたテキスト情報は日本経済新聞である．経済紙であるため，経済の変動を決定する要因が掲載されていると考えられる．

まず地方面を除く全記事に対して ChaSen[ChaS]を用いて形態素解析を行い，動詞・名詞・形容詞を抽出する．そして同一文中に隣接して出現した単語のうち，少なくとも一方に日経シソーラス[　日経]に収録されている経済専門用語が含まれる組合せのみを数え上げる．日経シソーラスとは日本経済デジタルメディアが公開している新聞記事検索のための辞書であり，約 1 万 3 千語が収録されている．単語ごとの出現頻度を数え上げるよりも，隣接した共起関係の出現頻度を数え上げることで経済動向に関する情報をうまく抽出でき，解釈も容易になると考えられる．従来の CPR 法では KeyGraph アルゴリズム[大澤 2006]を用いて重要語を求めているが，本研究は新聞記事を用いた長期予測であるため，日経シソーラスの語との共起をとることで網羅的な情報を抽出した．この数え上げを 1 ヶ月間の記事で繰り返し行い，閾値以上なら出現，閾値未満なら不出現とし，出現パターンを定義した．なお，閾値は 30 とした．1 ヶ月間に出現した共起関係のうち少なくとも一方に日経シソーラスの単語を含むものは約 10 万組存在し，閾値以上の共起関係の組合せは 400 から 500 組であった．

主成分分析

2.1 節の共起解析を過去 3 年間（36 ヶ月）の新聞記事に対して行い，閾値以上ならば 1，閾値未満ならば 0 とし，1 ヶ月ごとの単語の出現パターンを結合した行列を作成する．この時，訓練期間で少なくとも一回は出現した単語数は約 2 千語であり，36 行 2000 列の行列が作成されたことになる．この行列に対して主成分分析を行い，各月 15 個の主成分で記事を評価した．すなわち，1 ヶ月間の新聞記事を 15 次元のベクトルで評価し，そのベクトルを結合することで新聞記事の特徴量の時系列データが得られたことになる．

回帰分析

株価データは NOMURA400 の 19 業種，さらに日経平均，TOPIX を用いた．NOMURA400 とは，野村證券金融工学研究センターが提供しているデータであり，

日本株式市場の全銘柄から選定された市場代表性の高い投資ユニバースである．構成銘柄は日本株式市場の全銘柄の中から，アナリストの意見を基に選定された市場代表性の高い 400 銘柄である[野村]．そのうち，化学，鉄鋼・非鉄，機械，自動車，電機・精密，医療・ヘルスケア，食品，家庭用品，商社，小売り，サービス，ソフトウェア，メディア，通信，通信建設，住宅・不動産，運輸，公益，金融の 19 業種が対象である．日本の市場の分析には適している指標だと判断した．また，市場全体の動きを把握するための指標として日経平均，TOPIX も予測指標に加えた．指標iの時刻tにおける株価を$p_{i,t}$とすると，単位期間Δt(1 ヶ月，2 ヶ月，3 ヶ月)でのリターン$r_{i,t}$は下式で定義できる．

$$r_{i,t}=\frac{p_{i,t+\Delta t}-p_{i,t}}{p_{i,t}}$$

過去 3 年間の新聞記事から得られた主成分スコアと株価データを用いて次の回帰式を推定する．

$$r_{i,t}=a_{i,0}+\sum_{k=1}^{15}a_{i,k}x_{k,t}$$

$x_{k,t}$とは時刻tにおける第k主成分の主成分スコアである．回帰分析の際，AIC 基準[Akaike 74]におけるステップワイズ選択を行い，説明力の低い説明変数は回帰式に含めていない．

外挿予測結果

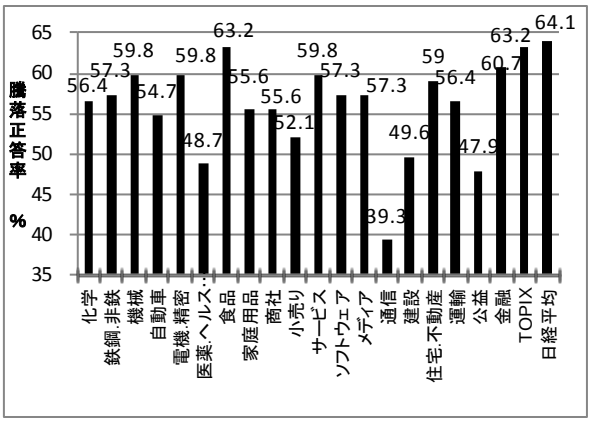
2 節の手法を用いて外挿予測の精度検証を行った．回帰式の推定および外挿予測の手順は以下の通りである．

1. 1998 年 1 月 1 日~2000 年 12 月 31 日の 3 年間(36 ヶ月)の新聞記事で説明変数の訓練データを作成する．
2. 各月の説明変数で翌月末の終値を被説明変数とする重回帰式を推定する．ただし，この時推定するのは 2.3 節で示したリターンの値である．
3. 推定された回帰式に直近のテキストデータ(2001 年 1 月 1 日~2001 年 1 月 31 日)から得られた主成分スコアを代入することで翌月末(2001 年 2 月末)の終値の推定を行う．
4. 訓練データの作成開始日を 1 ヶ月ずつ遅らせることで回帰式を毎月更新していき，2010 年 10 月末までの推定を行う．

1 ヶ月後の予測においては 9 年 9 ヶ月の 117 回の推定を行った．また，2 ヶ月後・3 ヶ月後の予測の場合，リターンにおける単位期間Δtを調整することで推定を行った．

1 ヶ月後の外挿予測結果

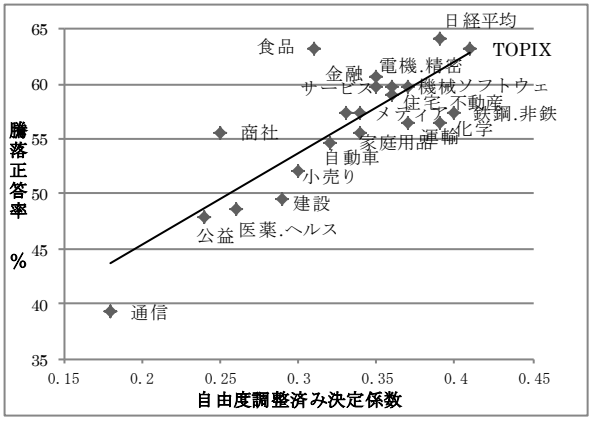
外挿予測の騰落正答率の結果を以下に示す．この結果は外挿予測を行った回数のうち，騰落が一致していた回数の割合を百分率で表したものである．



1 ヶ月後の予測の騰落正答率は市場全体の動向を表す TOPIX や日経平均といった市場平均株価で 63.7%であった．また，投資判断への適用可能性の目安となる 55%以上の正答率は 7 割以上の業種(予測指標 21 のうち，15 指標)で達成することができた．55%という騰落正答率は AI を用いた実際のフ

ァンドも目指している数値である[AERA]．本研究では日次よりも長期の月次予測で，しかも約 10 年間という長期間の外挿予測テストで安定した結果を示している．これは投資家から要望の高い長期予測の精度を大幅に改善したということである．

1 ヶ月後の予測においては，訓練期間の重回帰分析の自由度調整済み決定と外挿期間の騰落正答率に正の相関が見られた．その様子を以下に示す．図中の直線は一次の近似曲線である．



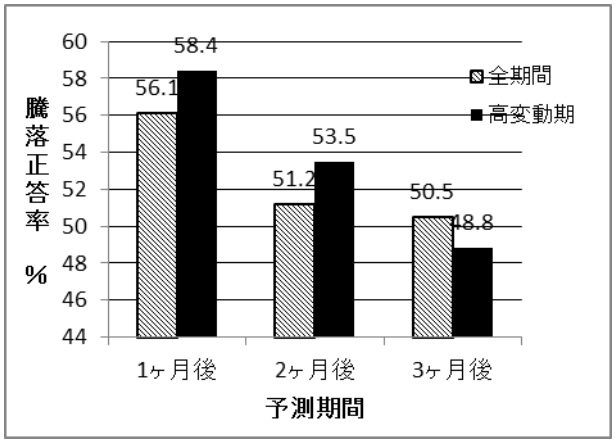
自由度調整済み決定係数とは回帰式の当てはまり度を表す指標であり，回帰式が過去の変動をよく説明できているほど騰落正答率が高くなる傾向が見られた．これは訓練期間で当てはまりのよい回帰式を作成できているほど外挿予測の精度も高まるということであり，内挿の段階で外挿予測の精度を大まかに推し量ることができる．また，図 2 からも明らかなように，本研究では過剰適合が生じていない．過剰適合とは，訓練期間で当てはまりの高いデータを作成するあまりに汎用性に欠け，未来のデータの予測には適さない形となってしまうことである．新聞という経済動向に関する網羅的な情報を扱うテキストデータから，主成分分析によって合成変数を作成し，さらに回帰分析の際にステップワイズ選択を用いることで過剰適合が回避されたと考えられる．

一方で，通信や公益や建設，医療・ヘルスケアに関しては自由度調整済み決定係数，外挿予測の騰落正答率ともに低い水準であった．これらの産業が内需産業と呼ばれ，新聞記事からこれらの産業に関する特徴量がうまく抽出できなかったためであると考えられる．

実際に抽出された共起単語のペアを見てみると，「長期金利 - 上昇」や「デフレ - 克服」などといった経済全体に作用すると考えられる共起，「米国 - トヨタ自動車」や「住宅 - 融資」のように個別の業種に影響を及ぼすと考えられる共起が見られた．これらの出現パターンを用いて回帰式を推定しているため，特徴量がうまく抽出できる業種とそうでない業種が存在する．それが図 2 の自由度調整済み決定係数に表れている．しかし，テキストという人間が解釈を行うことのできる情報を用いて株価予測を行っているため，各々の推定に対してどのような情報が実際に効果を持っていたのかを視覚的に確かめることができる．これは過去の変動に対して分析が可能である事，そして未来の株価の推定の際に用いることのできる可能性を示している．

予測期間別・高変動期の予測正答率

前節では 1 ヶ月後の外挿予測について述べたが，本手法によってどの程度先の未来まで推定できるのかを検証した．期間別の外挿予測の騰落正答率を下図左の斜線棒グラフで示す．期間別とは予測期間の長さであり，2 節で述べたように 1 ヶ月後，2 ヶ月後，3 ヶ月後の 3 つの期間について予測精度を検証した．



予測期間が長くなるにつれ，その間に起こる事象の数が増えるためテキストと経済変動の相関は弱まる．回帰式に用いられた主成分の見てみると 2 ヶ月後・3 ヶ月後の予測では，1 ヶ月後の予測の 1.5 倍から 2.0 倍の数の主成分が用いられていた．これは説明力の弱い説明変数を多く用いることで

経済変動に合致する回帰式を作成していたことになり，過去の変動の説明はできても未来の変動に対する予測力が落ちる．実際に，期間が長くなるにつれて予測精度が悪くなる傾向が見られた．逆に，1 ヶ月後の予測に関して 2 ヶ月後，3 ヶ月後の予測よりも少ない主成分による推定で好成績を収めたということは，それだけ経済の変動に直結した主成分を作成できていたということである．

予測期間が長くなるにつれて騰落正答率が下がる一方で，変動の大きな時期における予測正答率では 1 ヶ月後・2 ヶ月後の予測で 2 ポイント以上の改善が見られた(図 3 右の黒棒グラフ)．このことから本手法は 2 ヶ月後の予測まで有効であるということができる．なお，高変動期の予測とは下式によって定まる，時刻 T から始まる訓練期間 36 ヶ月におけるリターンの標準偏差 σ の 0.1 倍よりも絶対値が大きな変動を推定した場合のみ予測を行ったものである．

$$\sigma^2 = \sum_{t=T}^{T+35} (r_{it} - \overline{r_{it}})^2$$

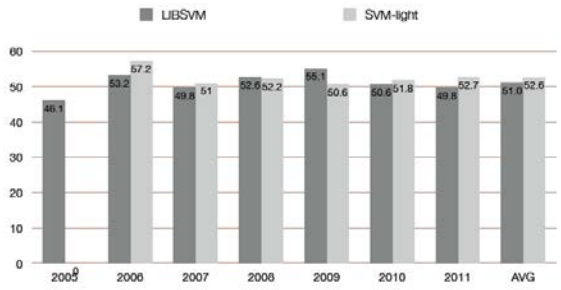
この閾値によって定めた高変動期は全予測期間のおよそ 3 分の 1 であった．予測した回数のうち，騰落が一致した割合を図中の黒い棒グラフで示している．

(2)様々な分類手法の精度比較

ここでは、新聞記事を用いた騰落予測に関して、教師付き学習手法として広く用いられる SVM に特に着目し、その精度向上に関する様々な手法の効果を調査した。

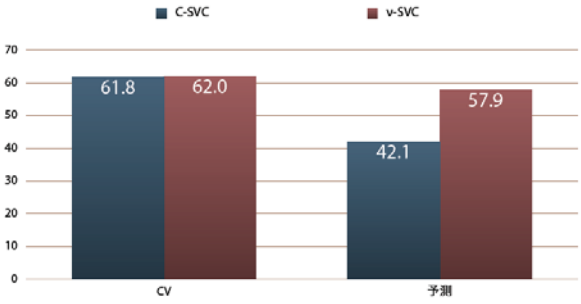
・SVM 分類器に関して、パラメータ調整を網羅的に行うことの効果

SVM 分類器のライブラリ LIBSVM におけるパラメータ調整機能を利用することで、分類精度を向上させることができるか否かの検討を行った。(下図、各年ごとの正解率) しかしながら、結果として，2005 年から 2011 年の 7 年分の予測においては、平均精度はむしろ低下した。



・ν-SVM を用いることによる予測精度の安定化

SVM の定式化の一種であるν-SVM において、最適なパラメータνの推定を同様に行うことによる、予測精度の変化についても考察を行った。2011 年 1 月のデータに対し予測を行ったところ（下図、CV：訓練データの予測精度、予測：テストデータの予測精度)、訓練データに対する精度はν-SVM を用いない場合と同等であったのに対し、テストデータに対する精度が向上することを確認した。これにより、ν-SVM を用いパラメータνの推定を適切に行うことにより、予測精度の安定化を図れることが期待できる。



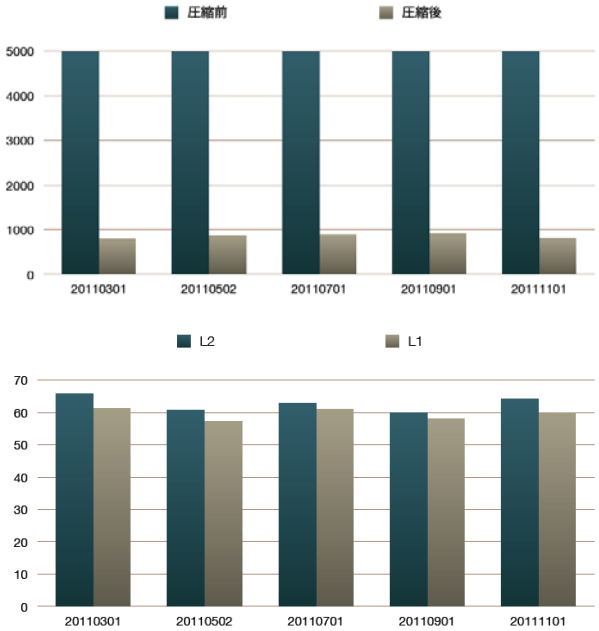
(3)適切な素性選択法に関する研究

大規模データに対する機械学習手法の適用の際、データの特徴量を削減することで、より高速・省メモリの処理を実現することができる。そこで、次元削減手法として広く用いられる L1 正則化の効果を確かめた。また、分類器とは独立に、値段を上昇させやすい単語、下落させやすい単語の抽出を行うことによって、これを特徴量として用いるという次元削減手法が考えられる。この手法についても検討を行った。

(3.1)L1 正則化を用いた特徴選択による効果の調

査

特徴量削減効果のある L1 正則化を、通常の L2 正則化と比べた結果を下図に示す。次元削減の効果に関しては、元の素性を 5,000 個とした場合、素性数を 1,000 個弱に減少させることができた。しかしながら、予測精度を僅かであるが悪化させることが判明した。



(3.2)ベイズ的スムージングを用いた素性選択

素性選択の手法として、より直接的に単語と価格の関係を用いた手法の検討を行った。ここで仮定する入力は、文書とそれに付随する実数値であり、この実数値を本稿では便宜的に「価格」と呼ぶ。目的は、より「価格」に密接に関連した単語を取得することである。取得結果は、機械学習の素性として用いることで、少ない次元で効果的な予測が行えることが期待できるほか、経済分野における極性（価格）付きの辞書として、重要文書を人手で絞り込む際の検索語として用いる等の応用が考えられる。このため、入力された各文は、あらかじめ形態素解析され、単語のリストに変換されている。

「単語の価格」を求める最もナイーブな手法としては、各単語について、それが登場する文すべての価格を集め、その平均を計算することである。しかしながら、単純に平均値を用いた場合、低頻度語に関する評価値の信頼性が低くなるという問題がある。例えば一回しか登場しない単語について、たまたまそれが登場する文書の価格が高い値だった場合、その一回の登場のみで、単語の評価値が高くなってしまう。この問題を解消するため、ベイズ統計に基づくスムージング手法を用いる。これは、価格の生成に関して何らかの分布を仮定し、そのパラメータの事後分布を計算しその平均を予測値とする手法であり、データへのオーバーフィッテイングを解消する効果がある。

(3.2.1)ベイズ的手法による評価値のスムージング

ベイズ的手法のスムージングを行う際、最も単純な手法として、価格 price(d)を正規分布から生成された値と仮定する手法がある。これについて、パラメータを、頻度が高い単語を高く評価するよう調整した結果が以下である。

正規分布を仮定した場合の高価格単語：
切り離し(1.11), カーライル(1.03), 囲い込み(1.03), ＴＯＢ(1.03), 相乗(1.03), 気配(1.03), 付か(1.02), 分割(1.02), 配当性向(1.02), 上方(1.02), ＡＯＫＩ(1.02), 初値(1.02), ウォルマート(1.02), ＭＢＯ(1.02), ベラ(1.02), 有力(1.02), 非公開(1.02), 〃(1.02), 大型(1.02), 仕入れ(1.02), 復配(1.02), 規制(1.02), ストップ高(1.02), 増配(1.02), エムビーエス(1.02), 黒字(1.02), 家主(1.01), 倍(1.01),

正規分布を仮定した場合の低価格単語：
更生(0.86), 民事(0.89), 破たん(0.90), 申請(0.92), 窓口(0.92), 法(0.93), 値幅(0.94), 国民(0.94), 撤廃(0.94), 相談(0.94), 不能(0.95), 取り立て(0.95), 公庫(0.95), 余波(0.95), 恐れ(0.95), 破産(0.96), 負債(0.96), 除外(0.96), 代用(0.96), 債権(0.96), 制限(0.96), 呼び値(0.96), 手続き(0.97), 注視(0.97), ジワリ(0.97), 再生(0.97), 波紋(0.97), 広がる(0.97), 貸出(0.97), 引き当て(0.97), 呼値(0.97), 取り消し(0.97), ショック(0.98), 先(0.98), 不透明(0.98), (0.99), 解雇(0.99), 不振(0.99), 粉飾(0.99), 直撃(0.99), 地銀(0.99), 記載(0.99),

さらに、正規分布ではなく、二項分布を仮定することもできる。これは、価格の実数値をそのまま利用するのではなく、上昇（1）か下落(0)の２値に変換して判定する手法である。価格が 1.01 以上のものを上昇、0.99 以下のものを下落と定義した場合の結果を以下に示す。

二項分布を仮定した場合の高価格単語：
上方(1.8107), 増配(1.8000), 設定(1.5989), 枠(1.5806), 倍(1.5050), 取得(1.4838), 黒字(1.4543), 寄与(1.4318), 買い(1.4293), ＴＯＢ(1.4208), 好調(1.3696), 分割(1.3500), 筆頭(1.3211), 分解(1.3115), 来期(1.2974), 増(1.2963), がん(1.2895), 短縮(1.2595), ストップ高(1.2587), 伸びる(1.2582), 優先(1.2500), 増益(1.2478), 分野(1.2477), 単位(1.2423), 承認(1.2414), 発表(1.2357), 量産(1.2269), 復配(1.2239), 受託(1.2210), 増産(1.2208), 画像(1.2189), 炭素(1.2164), 買収(1.2120), 剤(1.2112), 資本(1.2086), 他(1.2082), 基板(1.2047), 株式(1.2038), 狙う(1.2035), . (1.2020), 実力(1.2016), 貢献(1.2013),

二項分布を仮定した場合の低価格単語：
[分売(0.4596), 立会(0.4675), 下方(0.4778), 外(0.5468), 赤字(0.5501), 不振(0.6238), 減(0.6288), 無配(0.6630), 申請(0.6993), 売り出し(0.7110), 響く(0.7159), 減益(0.7272), 不能(0.7388), 見送り(0.7393), 勧告(0.7443), 証(0.7471), 響き(0.7500), 最終(0.7541), 一転(0.7577), 民事(0.7688), 回収(0.7718), 恐れ(0.7778), 捜索(0.7792), 法(0.7824), 有価(0.7844), 転落(0.7861), 遅れ(0.7865), 損(0.7865), 損失(0.7866), 減少(0.7884), 落ち込む(0.7963), 破産(0.7988), 遅延(0.8000), ストップ安(0.8000), 認定(0.8026), 本店(0.8032), 信用銘柄(0.8039), 取り消し(0.8045), 選定(0.8070), 停止(0.8079),

両者を比較すると、正規分布を仮定した場合、「破たん」「更生」といった、極端な下落を伴う単語が多く抽出されているのに対し、二項分布では、「赤字」「下落」といった、それほど極端ではないが一般的な単語を抽出することに役立つと考えられる。

(3.2.2)sLDA による、価格を考慮した単語のクラスタリングの作成

単語を単位とした抽出手法では、例えば「会社更生法申請」が「会社」「更生」「法」「申請」と分解され、例えば「更生」と「申請」が共に低価格単語の上位に来ることになる。しかしながら、これらの単語はほぼ同一の文書に登場するため、リスト中に両者が別々に登場することは、例えば利用者が低価格の文書を取り出したい場合に利用する際、無駄が生じることが予想される。この問題を解消するため、近年テキストマイニングの分野で広く用いられる「トピックモデル」の適用を試みた。具体的には、教師付き潜在ディリクレ配分法（sLDA）と呼ばれる手法を適用し、単語クラスタと、そのクラスタの価格を出力させる。建設関連銘柄に関する記事について、手法を適用した結果が以下である。

高価格クラスタ（建設関連銘柄）：
株(19.0000), 買い(18.0000), 発表(13.0000), ほか(10.0000), 状況(6.0000), 他(3.0000), 窓口(1.0000), 改善(1.0000), 相談(1.0000), 関連(1.0000), 報告(1.0000), 開設(1.0000), 化学(1.0000), 喚起(1.0000), 社(1.0000), 受注(1.0000)

低価格クラスタ（建設関連銘柄）
破たん(26.0000), 申請(19.0000), 法(18.0000), 県(15.0000), 更生(11.0000), 影響(10.0000), 再生(10.0000), 社長(9.0000), 先(9.0000), 相談(8.0000), 企業(7.0000), 上(7.0000), 者(7.0000), 発注(7.0000), 株主(7.0000), 訴訟(7.0000)…

最大の価格を持つクラスタが、自社買い発表に関する単語であるのに対し、最小の価格を持つクラスタは会社更生法申請に関するものであり、sLDA の有用性がある程度確認できた。

しかしながら、例えば「黒字」と「赤字」が同一のクラスタに入ってしまう等、価格を反映しない単語クラスタが形成されるリスクもあることが確認できた。

(3.2.3)用例抽出アルゴリズムを応用した、長い文字列の抽出

上述の sLDA では、評価値を考慮しつつ単語クラスタを生成することが難しいため、より直接的な方法として、単語ではなく、任意の長さの文字列を対象とし、「コーパスの特徴を俯瞰できるような文字列」を抽出し、利用者が「ポジティブな表現」「ネガティブな表現」について直感を得る手助けとするためのアルゴリズムの開発を行った。具体的には、用例抽出アルゴリズム Kiwi[Tanaka 2005]を参考に、「文字列の長さ×評価値の和」で各文字列をスコア付け及びランキングする手法を開発した。さらに、経済ニュースには数値情報を含む文字列が多いため、テキスト中の数値表現に対する処理も加えた。評価値は高い場合にプラス、低い場合にマイナスとなるように補正し、このため、「長い文字列で、頻度が高く、上昇傾向にある文字列ほど上位」「長い文字列で、頻度が高く、下落傾向にある文字列ほど下位」となる。スコア上位・下位の抽出結果が以下である。

（ただし、固有名詞は＊で置き換えてある。）

評価値合計を利用した場合の高価格文字列：
[（＊株取得枠設定），、＊，＊、，（ピックアップ）＊,[万~５０兆]，＊（＊，上方修正，、他（＊，へのＴＯＢ。，増，黒字，ホールディングス，ス、，月，ポイント，ット，業，高，イオン，＊に，好調。，クライマックスシリーズ??，ンティア，益，＊氏（新社長）＊（会社人事），株の貸借取引で申込停止措置。＊，倍，予想を上，、前期，子会社化，＊へのＴＯＢ，＊ロップフェニックス??，向け，...

評価値合計を利用した場合の低価格文字列：
[銘柄の選定取り消し。，生法申請，株の制度信用銘柄，最終赤字，、純利益[１~９９]%減，（立会外分売），赤字[１００万~９．６兆]円，更生法，法申請??，金、＊株の貸借取引，株を日々公表銘柄に指定。，申請、負債[１２億~２．７兆]円。，法申請、，、ＡＤＲ活用、親会社と来年[５~５]月合併，債権、回収不能の恐れ。，ホールディングス（ストックオプション），純利益[１~９９]%減、，株の貸借取引で注意喚起。，パシフィック，＊証金、＊株の貸借，（会社人事），ジョイント，、他（インフォメーション），。＊，下方修正，制度信用銘柄および貸借銘柄，

この結果を見ることで、「上方」「申請」「取り消し」等が、より長い文字列の一部となっていることがある程度確認できる。しかしながら、文字列どうしの重複がいくつか見られており、これらの抽出結果に関して、まだ加工の余地があることが観察された。

また、文字列抽出に関しても、単語抽出と同様、「1.01 以上をプラス」「0.99 以下をマイナス」として、絶対値ではなく上昇・下落の出現回数で評価するスコア関数を用い、同様に上位・下位文字列の抽出を行った。この結果が以下である。

二値化された評価値合計を利用した場合の高価格文字列:
[（＊株取得枠設定），、，へのＴＯＢ。,[万~５０兆]，ー，ス，株の貸借取引で申込停止措置。，上方修正，＊（＊，＊へのＴＯＢ，に，最終黒字，＊＊，完全子会社，「年明けうどん」，ン，うどん店，子会社化，＊（会社人事），純利益，。「＊賞与」政権揺さぶる，の貸借取引で申込停止措置。＊，業，ネット，月，から，株の貸借取引で申し込み停止措置。，増，予想を上，ル，イ，ＴＯＢ。＊，新興??,...

二値化された評価値合計を利用した場合の低価格文字列:
[　（会社人事），、純利益[１~９９]%減，（立会外分売），最終赤字，赤字[１００万~９．６兆]円，ブランド，純利益[１~９９]%減、，株を日々公表銘柄に指定。,??＊（ＮｅｗＦａｃｅ），下方修正，益[６~９８]%減、今期，＊＊，氏（新社長）,[９００万~５．７兆]円、[１~１２]?[２~１２]月，決算から??[１~１１]?[１~１２]月、＊、，字[４００万~９．６兆]円、，（フォローアップ），株の日々公表銘柄指定を解除。，。＊，金、＊株の貸借取引，ス（会社，期、純利益[１~９９]%，益[２~９９]%減（業績ダイジェスト），＊社長＊，株の信用取引，株の貸借取引で注意喚起。，...

二値化されたことにより、単語の場合と同様、決算関連の文字列が上位に来ていることが観察できた。一方、特に、上昇文字列に関して、二値化前と後で、単語の場合程の違いが見られなかったため、これらをより差異化できるようなスコア関数についても検討の余地がある。

参考文献

[AERA] AERA’2012.2.13, pp.62, 朝日新聞出版.

[Akaike 74] Akaike, H.: A new lool at the statistical model identification, IEEE Transactions on Automatic Control, Vol.19, pp.716-723 (1974).

[ChaS] ChaSen ホームページ : http://chasen.naist.jp/hiki/chasen .

[高穂 2002] 高穂洋, 荒井隆行, 大竹敢, 田中衛 : ニューラルネットワークによる時期の株価予測－株価予測におけるフィルタリングによる特徴量抽出－, 電子情報通信学会技術研究報(NPL), Vol.102, No.432, pp.13-16 (2002) .

[伊庭 2006] 伊庭斉志 : 進化論的手法を用いた金融データの予測, 日本信頼性学会誌, Vol.28, No.7, pp.471-480 (2006) .

[和泉 2011] 和泉潔, 後藤卓, 松井藤五郎 : 経済テキスト情報を用いた長期的な市場動向推定, 情報処理学会誌, Vol.52, No.12, pp.3309-3315 (2011) .

[大澤 2006] 大澤幸生 : チャンス発見のデータ分析 - モデル化+可視化+コミュニケーション→シナリオ創発, 東京電機大学出版局 (2006).

[日経] 日経シソーラス, 日本経済新聞デジタルメディア http://vip-test2.nikkei.co.jp/help/contract/price/02/help_KIJI_thes.html

[野村] 野村證券金融工学研究センターホームページ : http://qr.nomura.co.jp/jp/n40/index.html※※

[Tanaka 2005] Kumiko Tanaka-Ishii, Hiroshi Nakagawa, ″A Multilingual Usage Consultation Tool based on Internet Searching ---More than

search engine, Less than QA″, The 14th International World Wide Web Conference (WWW2005) pp.363-371. 2005.

(2) 当初計画の達成状況について

本研究課題の目的は、従来研究してきた株価予測の手法を、言語資源の利用やより進んだ機械学習手法の利用を用いて高度化し、大規模データの処理に耐えうる手法に改善すること、また、手法の改善により、精度向上に繋げられるか否かを確認することであった。機械学習手法については、特にパラメータ調整に関して検討を行い、また、速度やメモリ効率向上に繋がる素性選択について、その可能性を検討した。高価格単語・低価格単語の抽出に関して、テキストマイニングの手法を応用し効果的な手法を開発することができたが、この結果を素性選択に活用できるかは未知数であり、確認の必要がある。言語資源利用については、経済辞書（日経シソーラス）を利用した手法の改善について研究を行った。大規模データ処理のためにオンライン学習を行う研究については、今年度内に実現することができなかったため、今後の課題としたい。

4. 今後の展望

テキストからの株価の予測を目的として、様々な手法の検討を行った。ここで得られた知見をもとに、今年度達成できなかった大規模データ処理への適用を試みていきたい。また、そのために、特に素性選択の手法について、高速化や洗練化が必要であるため、これに関する研究も継続していきたい。

5. 研究成果リスト

(1) 学術論文（投稿中のものは「投稿中」と明記）

藏本貴久, 和泉潔, 吉村忍, 石田智也, 中嶋啓浩, 松井藤五郎, 吉田稔, 中川裕志, 新聞記事のテキストマイニングによる長期市場動向の分析, 人工知能学会論文誌 Vol. 28, No. 3 , pp.291-296 (2013)

(2) 国際会議プロシーディングス

なし

(3) 国際会議発表

なし

(4) 国内会議発表

藏本 貴久、和泉 潔、吉村 忍、石田 智也、中嶋 啓浩、松井 藤五郎、吉田 稔、中川 裕志, 新聞記事のテキストマイニングによる長期市場動向の分析, 2012 年度人工知能学会全国大会（第 26 回）, 2012 年 6 月 15 日, 山口.

(5) その他（特許, プレス発表, 著書等）

なし