金融分野におけるビッグデータ分析

古関 聰 金山 博 坪井 祐太 平出 涼 千葉 立寛
（日本 IBM 東京基礎研究所） 米持 幸寿 野村 尚（日本 IBM）

概要 マーケットデータや SNS，ニュースなどの膨大な各種テキストを蓄積し，分析を行うことが容易になってきている．本論文では，テキスト解析，データマイニング，分散処理を相互に活用したビックデータ分析手法の応用方法について概観を行い，特に金融分野での分析事例を紹介する．各事例についての分析の狙いや効果について論じた後，蓄積されたデータからの情報取得方法や分析の適用手法等の具体的なプラクティスについての議論を試みる．

1. はじめに

ネットワークの高速化，広域化，モバイルデバイスや各種センサの爆発的な普及，ストレージの大容量化やクラウド技術の普及により，多様なデータの蓄積が急速に進んでいる．データが大規模化する主な要因として，センサのインストール台数が増えていることやデータの検知間隔が非常に細微化していること，また，スマートフォン等のモバイルデバイスのユーザ数や電子情報が集積されるサイト数の増加により，特にテキストの情報が大規模に生成・蓄積されていること等を挙げることができる．

こういった大規模なデータを蓄積・解析しビジネスに応用する機運は年々高まっており，Hadoop[1]を中心とする分散処理基盤が普及するとともに，基盤上でデータを効率的に解析する研究やプラクティスが進んでいる．

本論文では，ビックデータ分析の応用として特に金融業界におけるいくつかの事例を取り上げ，分析の目的や狙い，実際に得られた結果について論じるとともに，具体的なプラクティスをもとに今後取り組むべき課題についての議論を試みる．

2. ビッグデータ解析の背景

2.1 ビッグデータ分析への期待

ネットワークの高速化，広域化，モバイルデバイスや各種センサの爆発的な普及，ストレージの大容量化やクラウド技術の普及により，画像やテキスト等の非構造化データを含む多様なデータの蓄積が急速に進んでいる．

多様なセンサの配置が社会の隅々に広がり，非常に細かい時間でデータが生成されているとともに，センサデバイスは広域ネットワークに接続され，大規模な計算システム上にデータが蓄積されている．また，スマートフォン等の普及に伴い SNS 等の電子サイトに集積されるデータは爆発的な増加を見せている．各種メディア企業においてもコンテンツのデジタル化が進み，長期間に渡る大規模なデータの蓄積が可能になっている．

このような背景をもとに，いわゆる伝統的な構造化データに加え，画像やテキスト等の非構造化データを含めた多様なデータの大規模な蓄積が近年急速に進んでおり，今後もその傾向が続くと考えられる．ここで，非構造化データとは，固定されたデータスキームで必ずしも定義できない型を持つデータを指すものとする．一方で，そのような多様なデータを集積し縦横に分析を行うことで，これまでに得られなかった知見を得ようとする新しい試みが行われており，さまざまなビジネス領域において大規模データの解析への期待が高まっている．

このような解析のパターンとして，いくつかの類型があると考えられる．代表的なものとして，まず，主にテキスト等の非構造化データの大規模な集積を社会の全体的な動きの反映として捉え，それを解析することで人々の考えや感覚についての知見を得ようとするものが挙げられる．また一方で，ビジネス上関心のある事象についてモニタリングした数値データを対象とした既存のデータ解析をさらに進める目的で，上記の非構造化データを取り込むといったものが考えられる．

こういった非構造化データを含む大規模データの分析を実際に可能にする要素技術の発達が近年顕著であり，今後は広いビジネス領域での応用が見られると考えられる．

2.2 ビッグデータ分析を可能とする主要技術

本節では, 2.1 節で述べた分析を可能とするための主要技術について概観する．我々は，ビッグデータ解析のための主要技術として，テキスト解析技術，データマイニング技術，分散処理技術に注目している．図 1 は各技術とそれらを組み合わせた応用が考えられる領域について図示を試みたものである．

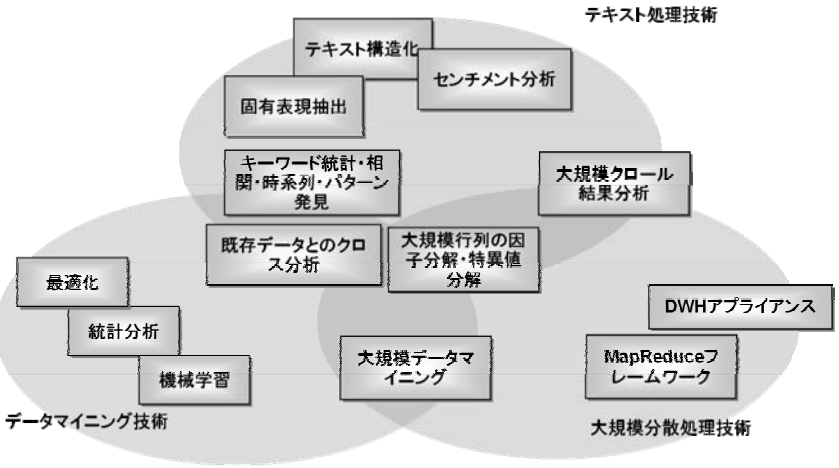

図 1. ビッグデータ分析のための主要技術

データマイニング技術はデータから有用な知識を抽出するためのコア技術であり，データの相関の発見，統計値の算出データの分類や予測，クラスタリングといった技術の総称をここでは指している．これらの技術は SPSS[2]や SAS[3], R[4]といったソフトウェアの充実により，解析アルゴリズムを応用することが容易になってきている．ビジネス上で生成・取得されるデータをこれらのソフトウェアで分析するといったプラクティスが一般に行われてきている．

テキスト解析技術は，テキスト等の非構造・準構造化データの蓄積が膨大であることを考慮すると非常に重要な技術である．ここでは，コンピュータによる解析という観点から，非構造化データを構造化データに変換するという意味でテキスト解析技術を取り上げる．テキストを構造化するためには，文字列から単語などの意味単位を抽出し，意味単位間の関係を構造的・文法的に解析する必要がある．こうして意味単位の上位の意味的解析が可能になる．そのために基本となる技術として，日本語では形態素の解析，英語では品詞のタグ付けが挙げられる．また，代表的な意味的解析としては，名前，地名，住所，商品名といった固有概念を抽出する固有表現抽出[5]や，Positive/Negative といった感情表現を意味単位の出現パターンより抽出するセンチメント分析[6]が挙げられる．

分散処理技術は，複数のコンピュータに処理を分散して処理を行う一般的な技術であるが，ここではビッグデータを対象とした技術として，配置した大規模データに対し並列処理を行う技術に注目する．コモディティ・コンピュータによりクラスタを構築し，クラスタ上に構成された分散ファイルシステムと，Map 処理，Reduce 処理を組み合わせて記述するフレームワークで分散処理を実現する Hadoop の成功により，大規模蓄積データの分散処理への注目が高まっている．IT 企業に限らず様々な企業が Hadoop 等を活用した大規模データ処理に取り組み始めている．

これらの主要技術は単独でも重要なものであるが，ビッグデータの解析は，これらの技術の組合せが重要である．

すなわち，非常に大規模なテキスト蓄積を構造化するための，テキスト処理技術と分散処理技術の組合せ，テキストを構造化して得られるデータを機械学習の入力とするようなテキスト技術とデータマイニング技術の組合せ，機械学習における学習アルゴリズムを並列化し大規模なデータを解析するデータマイニング技術と分散処理技術の組合せといった応用が考えられる．このような技術の組合せによる応用は，今後のビッグデータ解析の実現のための必須な研究や実践の領域になると考えられる．

3. 金融分野におけるビッグデータ分析

第 2 章で述べたとおり，ビッグデータ解析の代表的なパターンとして，テキスト等の非構造化データの大規模な集積を社会の全体的な動きの反映として捉え，それを解析することで人々の考えや感覚についての知見を得ようとするものがある．また，ビジネス上関心のある事象をモニタリングした数値データを対象とした既存のデータ解析をさらに進める目的で，上記の非構造化データを取り込むといったアプローチが考えられる．

ここでは，金融分野についてこのようなアプローチを具体化する．まずは，テキストから経済・金融についての人々の認識をどのように取り出すかという問題を議論する．たとえば，さまざまな報道データが電子化され，長期間にわたり企業や経済についての意見や記述が蓄積されている．このようなデータからいかに経済状況を把握するための重要な情報を分析・取得するかが重要であ

る．このような情報を大規模なデータ蓄積から効果的に抽出することができれば，金融ビジネスに携わる企業が市場を把握するために非常に有益であると考えられる．また，Twitter[7]などの SNS には，非常に膨大な数のユーザからの発信情報が蓄積されている．このような不特定多数からの非常に大規模な情報蓄積に対しては，経済状況に関連した情報をいかにして効率よく取得するかが重要となる．

また，金融領域においては，データマイニング手法を応用してさまざまな経済現象をモデル化し，その説明を試みたり現象を予測するといったことが数多く行われている．このような解析手法では，一般的にマーケットや各種経済統計から取得可能な数値データが入力となる．ビッグデータを活用した分析では，テキスト等の膨大な非構造化データから得られる情報を数値化し，数理解析の追加的な入力とすることが考えられる．このような入力がこれまで必ずしも精緻にモデル化できなかった経済現象の有力な手がかりとなることは十分に期待できると考えられる．

4. ビッグデータ分析事例

本章では，我々がこれまでに検証した分析事例を複数例とりあげ，それぞれの分析の目的や手法について紹介を行った後，具体的なプラクティスより得られた知見について論述する．

4.1 節では氾濫する情報からトピックに合致するものを選択する技術，4.2 節は数理解析の面から，4.3 節は言語処理の観点からその分析精度を高める技術である．

4.1 Twitter キーワード分析事例

4.1.1 分析の目的

本節では，SNS から経済の状況を把握する試みのひとつとして，Twitter から企業の活動に関連したテキストをどのように効果的に取得するかを実証した事例を紹介する．

Twitter には日々膨大な量の投稿が行われており，この中から適切な情報を抽出するのは容易でない．しかしながら，投稿された「つぶやき」の中には企業あるいはその製品の評価といった情報があり，それらを効果的に抽出・集計することができれば製品販売といった観点から企業の業績を理解する一助となると考えられる．このような情報は経済アナリストへのインプットとなるほか，他の事例で述べるようなデータマイニングへの入力データとしても活用できる．

4.1.2 情報抽出システム

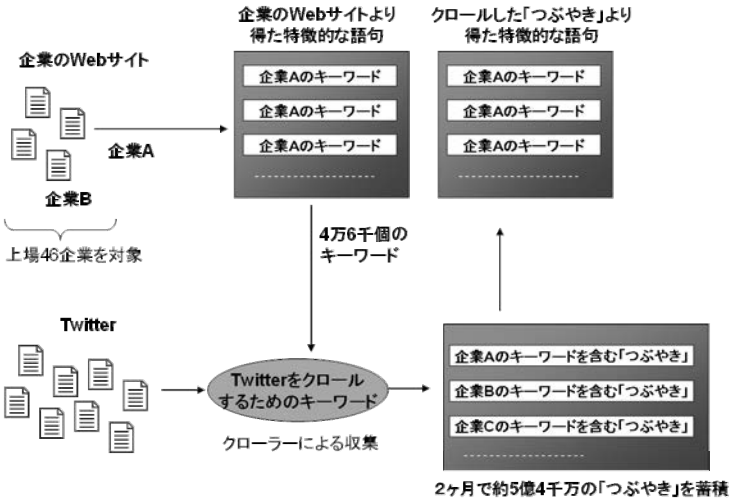

図 2. Twitter キーワード分析

図 2 に，このような情報抽出の仕組みを実現したシステム図を示す．本システムでは，企業に関係した情報をキーワードとその特徴量を手がかりに抽出・分析する．まず，企業の Web サイト（特に企業の営業実績に関わる情報が豊富な IR 情報ページ等）のテキストをクロールし，企業毎にその企業での特徴的なキーワードを抽出する．ここでキーワードの特徴量を，

企業のページのテキスト群におけるそのキーワードの出現比／全体のテキストにおけるそのキーワードの出現比

としている．このようなキーワードを初期セットとして Twitter 文書をフィルタリングする（キーワードによる Search API[8]を使う，あるいは Stream API[9]で取得した Tweet をキーワードで選別する）ことで，ターゲットの企業に関連した書き込みの文書セットを得ることができる．ただし Web ページから得たキーワードが書き込みの中でよく使われているとは限らない．そこで，キーワードの特徴量を，

その企業のキーワードが含まれている Twitter テキスト群におけるそのキーワードの出現比／全体の Twitter テキストにおけるそのキーワードの出現比

として再計算し，このようなプロセスを繰り返すことで，より適切なキーワードが得られるものと考えられる．

図 2 のシステムにおいては，上記のようなプロセスを非常に大規模なテキストセットに対しても適用できるよう，商用 Hadoop 製品である IBM InfoShpere BigInsights[10]と同テキストマイニング製品である IBM Content Analytics (ICA)[11]を組み合わせて実証システムを構築した．本実証では，対象企業を 46 個の上場企業と

し，約 46,000 個の初期キーワードをもとに Twitter サイトのクロールを行い，BigInsighgs 搭載のスクリプト言語 Jaql[12]から ICA の形態素解析を呼び出して上記の統計量等を 4 台の Hadoop クラスタを用いて計算した．本実験においては，この時点での Twitter の API の制限下で当クラスタのリソース下で最も効果的なクロールが行えるキーワード数として 1 企業あたりでの 1,000 個のキーワードを指定した．

4.1.3 分析結果と課題

本実証の分析結果の一部として，図 3 にキーワードの特徴量とトピック妥当性間の関係を示す．株式会社パルコ（業種：小売），株式会社 ADEKA（業種：化学）を対象として，キーワードの特徴量と，トピック妥当性について，無作為に抽出したサンプルに対してプロットした．ここで，横軸はキーワードの特徴量を示す．また，縦軸はトピックの妥当性として，そのキーワードが含まれる Twitter 文書がその企業の活動に関連している比率を示すものとする．

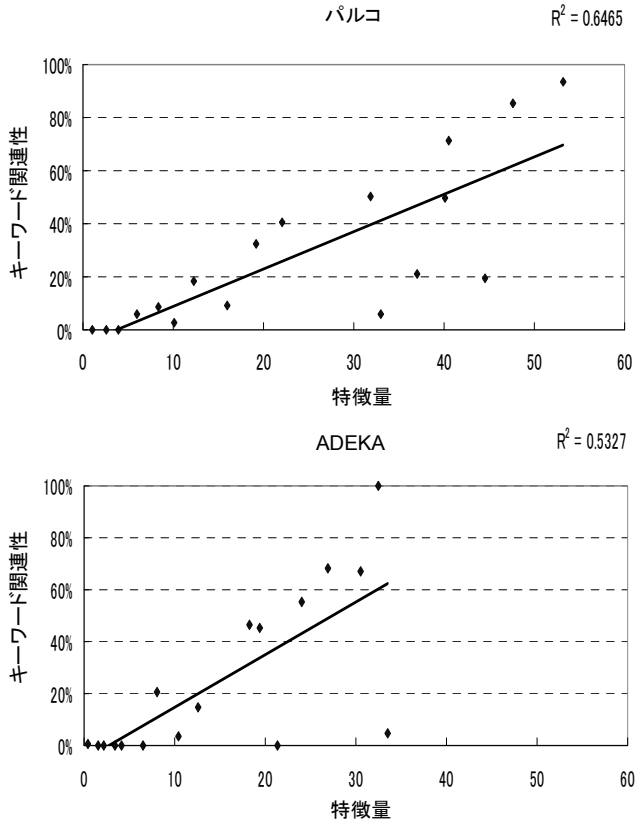

図 3. キーワードの特徴量とトピック妥当性の関係

図 3 の両チャートより，キーワードの特徴量とトピック妥当性には正の相関を読み取ることができる．このような傾向は今回の対象企業の中で，小売，メディア，化学，食品といった業種で顕著に現れた．一方で，例えば建設業，製鉄業，証券業といったその他の業種の企業では，高い決定係数を持つような相関結果は得られなかった．これは，Twitter に投稿される情報は日々の生活に身近なものが多いということが理由して考えられる．したがって，このような分析は B2B というよりは直接ユーザーと関わりのある B2C 企業，あるいは B2B であってもよりその製品やサービス内容がユーザの生活に関連している（例えば食品の重要な成分を提供している等）企業が向いている．そのような企業に関して，本実証のようなキーワードの特徴量を手がかりとして膨大な Twitter 文書の集約と分析を行う手法が有効であると考えられる．

4.2 経済指標予測事例

4.2.1 分析の目的

本節では，経済指標予測の取り組みのひとつとして，米国の ISM（Institute for Supply Management）製造業景況指数（以降，ISM 指数）の大規模な非構造データを活用し予測する統計モデル構築事例を紹介する．

ISM 指数は米 ISM が製造業 300 社以上の購買担当役員に「新規受注・生産・雇用・入荷遅延・在庫」の 5 項目について「増加，同じ，減少」のアンケートを行い，その回答をもとに算出される．50 を上回ると景気拡大，逆に 50 を割り込むと景気後退を示唆しているとされる．ファンダメンタルズと呼ばれる経済指標は経済が今後どうなるかを予想するものであるとともに，為替などの短期的な金融取引にも影響すると考えられている．特に，ISM 指数は毎月第 1 営業日に発表されることから，米国の経済指標の中では最も早く発表される景気転換の先行指標として重要視されている．そのため，ISM 指数を自動的に予測することができれば未来の経済状況に対する示唆を得ることができる．

4.2.2 回帰モデル

ISM 指数は，その発表前にアナリストなどによる予想が行われる．同様に発表前の ISM 指数を自動的に予測する状況を想定して，前月の ISM 指数に加え失業率や小売売上高などの他の経済指標を説明変数として ISM 指数の値を予測する自己回帰モデルを構築した．さらに，アンケート結果という特性を考慮して，回答者のセンチメントに影響するであろう株価や米国の報道情報も説明変数とした．

本分析では，非構造化データである報道情報から，テキストに含まれる単語や単語の組合せあるいは報道情報に含まれる各種情報を非構造特徴として抽出し，数値化

の対象とした．なお，形態素解析を用いて単語は名詞や動詞などの内容語に限定した．次に，これらの非構造特徴を ISM 指数の発表頻度に合わせて 1 カ月間の平均頻度または出現の有無を示す 2 値として数値化した．この際，ある閾値以上の月数出現した特徴のみを使用した．

最終的に非構造特徴と ISM 指標・その他の経済指標・株価を組み合わせて説明変数とした．なお，ISM 指標以外の経済指数や株価・新聞特徴は前月との差分または対数差分を特徴量とし，また株価は非構造特徴と同様に ISM 指数の発表頻度に合わせて 1 カ月間の平均頻度として数値化した．

説明変数の数は経済指標だけで数百次元，単語の組合せまで含めると百万次元超となる．一方，ISM 指標は月単位のイベントのため，過去に遡って次元数 d に見合うデータ数 n を集めることは困難である．この d>>n となる状況に対応するため，カーネル法とリッジ回帰を組み合わせたカーネルリッジ回帰を用いた（詳細については文献 13）等を参照されたい）．カーネル法を用いることでパラメータ推定に用いる最適化問題は d に依存しない計算量となり学習が高速になる．また，リッジ回帰を用いることで過去データへの過学習を防ぐことが可能になる．ただし，リッジ回帰では正則化パラメータを決定する必要がある．正則化パラメータの選択には訓練データを学習用と検証用に分割してモデルを推定し絶対誤差を計算し，10 回異なる分割をしたときの合計絶対誤差（モデル選択時誤差）が最も小さいパラメータを選択した．また，予備実験よりカーネル関数は線形カーネル関数を選択した．

4.2.3 分析結果と課題

予測モデルを評価するために，訓練データとして 1999 年 1 月から 2008 年 11 月までの n=119 カ月を使用し，評価用データとしては 2008 年 12 月から 2011 年 11 月までの 36 カ月を使用した．ISM 指標やその他の経済指標は最初の発表後にその値が改定されることがあるが，予測時に入手可能な値のみを用いた．株価を含む経済指標に基づく説明変数は 377 次元であった．報道情報は 1998 年 12 月から 2011 年 11 月までのデータを使用し非構造特徴を作成した．

表 1 に予測モデルの評価期間での平均絶対誤差，前月からの指数値の上下動の正解率（方向正解率），モデル選択時誤差を示す．方向正解率を評価尺度として併用したのは，ISM 指数の前月からの増減が景気に対する見方への変化を示しているため，金融市場への影響が大きいと考えたからである．

非構造特徴の説明変数は数多くの組合せを試したが，ここでは予測パフォーマンスが良好であった株価を含む経済指標のみと訓練データを使ったモデル選択時誤差が最も小さいモデル（経済指標＋非構造特徴）の結果を示す．また，比較対象として Bloomberg 社記事に掲載されたアナリスト予想の中央値を予測値としたときの評価結果も掲載した（市場予想）．

表 1 に示したとおり，訓練データで最も性能が高いと判断された，モデル選択時誤差が最も低い“経済指標+非構造特徴”は平均絶対誤差・方向正解率ともに“市場予想”より若干よい結果となった．このことから，過去データから学習した統計モデルが経済の専門家らと同等の予測性能を持つことが示された．

報道情報を説明変数に使った効果であるが，モデル選択時誤差は非構造特徴を使ったほうが誤差が下がっており，与えられたデータを説明するという観点では，非構造化データを取り入れたモデル構築の有効性が観測できている．しかしながら，予測パフォーマンスという観点においては，経済指標だけを用いたモデルの方が評価期間での性能は良かった．また，報道情報内の単語のみを特徴に用いたモデルは選択時誤差は比較的高く，モデル選択時には選ばれないという結果になった．これは，テキストデータからより高次の意味を取り出して活用したほうがモデルが改善できる可能性を示唆していると考えられる．

今後の課題としてはモデル選択手法の改善が挙げられる．モデル選択時では選ばれなかったが評価期間では良い性能を示す特徴集合があった．同様に，モデル選択時誤差を基準に経済指標の変数選択も試みたが，評価期間の性能向上には繋がらなかった．訓練データを増やすことや洗練されたモデル選択手法を用いることで，非構造化データを含む多くの説明変数を取り入れながら頑健なモデルを構築することが今後望まれる．

表 1．評価期間での市場予測と予測モデルの比較

	平均絶対誤差	方向正解率	モデル選択時誤差
市場予想	1.694	0.778	N/A
経済指標のみ	**1.591**	**0.806**	1.399
経済指標+非構造特徴	1.624	**0.806**	**1.391**

4.3 金融センチメント分析事例

4.3.1 分析の目的

本節では，非構造化データからの高次の意味内容を取

り出すプラクティスの一つとして，金融テキストのセンチメント分析事例を紹介する．

その目的は，金融商品や個別株式の取引の意思決定を補助するために，それらに影響を与えるような経済状況や会社の業績を直接的・間接的に示しているニュース記事を検出し，可視化することである．特定のアナリストによる分析には，主観が含まれたり，言及できる事象が限定されることが避けられない．一方で，最新の新聞記事など，より多くの情報源から重要な要素を自動的に抜き出すことによる客観性と網羅性により，個人の投資活動が活性化することが期待される．また，前節において，高次の意味内容を反映した入力変数が予測モデル構築の改善につながることが示唆されており，一次非構造化データの加工手法としての応用が期待できる．

4.3.2 センチメント表現抽出

市況のタイムリーな判断に有用な情報の一つに，対象ごとのセンチメント表現，すなわち positive，negative な内容がある．テキスト情報の中に現れるこれらの表現の多寡やその内容を用いて，好材料や悪材料を早期に発見する，市況の予測の指標や実際の株価の変動等との関連を分析する，といった応用が考えられる．ここでは，センチメント表現を高精度で抽出し，知識発見に役立てるための方法論について記述する．

Positive，negative な表現を抽出する技術はこれまでに多数提案されており，単なるキーワードの抽出に留まらず，構文解析の結果をもとに，否定などのモダリティ表現を捉えた高精度な分析をする手法などがある[6].

この抽出ができれば，特定の企業名を含む記事とともに現れる positive，negative の表現の数や増減を時系列で比較することができる．図 4 は製造業の特定の企業に関する例で，上段が positive，下段が negative の表現のグラフであり，色の濃さで増加量を把握できる．これが直ちに株価と連動するところまでは検証できていないが，どの時期に良いニュース，悪いニュースがあったか，その内容や原因が何であったかを効率よく調べることができる．

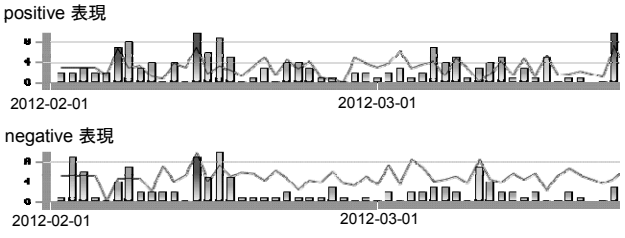

図 4. Positive，negative 表現の増減

特定の企業のセンチメントを分析する際の問題点として，一つのニュース記事の中に複数のトピックが述べられる場合があるという問題がある．A 社についてのセンチメントを調べる際に，A 社のキーワードと positive，negative が同時に現れる記事の数を観測すると，同じ記事であっても別の文脈の事象が多数同定されてしまう．この問題の解決のため，今後記事中のトピックを適切に自動分割する手法を開発していく必要がある．

4.3.3 センチメントの語彙の拡張

Positive，negative の極性を持つ語彙の知識として，「好きだ」「素晴らしい」または「嫌う」「悪い」といった直接的な評価を表す表現が既に用意されている．しかし，分野や状況に特化した語句がしばしば有用な示唆を与える．例えば，金融の分野では「上方修正する」「伸び悩む」といった述語のほか，「設備投資が…増える」「在庫が…だぶつく」といった述語句で表現されるものがある．

しかし，これらの分野別の辞書を逐一人手で作成すると高いコストがかかる．そこで，大量のテキストと分野非依存の辞書をもとにして，分野に特化した語彙の拡張を自動的に行う方法[14]を適用した．これは，positive や negative な表現が現れる文脈が存在する性質を利用して，既存の語彙から推定されるセンチメントの前後に出現する語句のうち，positive ないし negative の文脈に偏って現れるものを統計的に抽出し，新たなセンチメントの語彙を構築するという手法である．この手法は，他にも経済向けの研究に適用された実績がある[15].

本実証においては，経済関連のニュース記事を用いた語彙の獲得を試みた．表 2 に得られた語彙数，表 3 に自動獲得された語彙の例を示す．表 2 において，極性語とは「底堅い」「行き詰まる」のように動詞や形容詞単独で positive・negative の極性を持つ語句，極性句とは「知名度が…高い」「減益を…もたらす」のように，述語の格要素を伴った状態で極性を持つ語句として，それぞれ判定されたものの数である．「正解率」は，それらの極性語・極性句をそれぞれ 20 個サンプリングしたうえで，当該分野におけるセンチメントの語彙として妥当であるかを評価し，それらが出現する頻度による重み付けをした値である．

実験結果より，語彙の獲得数は情報源の量に比例していることがわかる．特に，ニュース記事より多くの極性語句を得られており，さらに大規模な入力を対象とすることでセンチメント分析の正確さが向上するものと考えられる．一方で，情報源の量が増えるに従い正解率が下

がっている．これは，経済と関連のない記事などによるノイズが大きいことによる．本結果は，製品やサービスの口コミを対象とした先行研究の実験[6]に比して正解率が低く，当該分野の困難さがわかるが，入力記事のエリアをあらかじめ選別，あるいは自動判別するような手法と組み合わせることでより実用性が向上することが期待できる．

表 2．自動獲得された語彙の数とその正解率

情報源	極性語	極性句	合計	正解率
金融月報 4 年分（48 記事・3,350 文）	8	63	71	84.5%
ニュース記事 1 カ月分（12,891 記事）	55	94	149	78.4%
ニュース記事 2.5 カ月分（40,124 記事）	107	429	536	77.1%

表 3．自動獲得された語彙の例

positive	negative
寄与する・しっかりする・有効求人倍率が…増加する・コストを…抑える・復興需要が…ある	冷え込む・炎上する・相場変動に…左右する・収益性が…低下する・特別損失に…計上する

本手法により自動獲得された表現は，辞書としてセンチメントの抽出の際に使われ，抽出される情報の量的な面はもちろん，そして具体性を増すという質的な面に寄与する．さらに，positive, negative の文脈に偏って出現する語句自体がしばしば興味深い．情報源とするテキストのトピックないし時期を絞り込んで情報抽出を実行すれば，好材料・悪材料となる主概念を容易に発見することができる．このような抽出結果は，経済の状態の効果的な概観を可能にする．また，語彙の出現頻度を大規模なデータから獲得し，4.2 節のような数理解析の入力変数とすることも有効であると考えられる．

5. おわりに

本論文では，ビッグデータ分析の金融業界における応用として，特に大規模非構造化データをどのように分析するかという観点から三つの具体的なプラクティスをとりあげ，分析の目的と分析結果あるいはその過程から得られた知見について議論を試みた．

本論文で述べたように，ビッグデータの主要な蓄積はテキストを中心とした非構造化データであることから，特にこのようなデータの分析例として，Twitter から経済現象としての企業活動を把握するための情報を効果的に獲得するためのキーワード分析，数理的な経済予測モデルへのテキストデータの取り込みの試み，及び，テキストからより高度な経済・金融記述を抽出するためのセンチメント分析を取り上げ，その手法と知見を紹介した．

このような大規模な非構造化データの分析は萌芽的な側面があり，必ずしもビジネスにすぐに応用できる段階に至っていない点もあるが，本論文の冒頭で述べたように，テキスト処理技術，データマイニング技術，分散処理技術を組み合わせた領域における技術の発展は，本実証結果を含めたさまざまな知見を取り込み今後も進んでいくと考えられる．当研究所においても，この分野の基礎的・応用的手法を継続して試みている．

参考文献

1) Apache, Hadoop 公式サイト, http://hadoop.apache.org/ （2012 年 11 月 26 日現在）.
2) IBM, SPSS 製品サイト, http://www-06.ibm.com/software/jp/analytics/spss/ （2012 年 11 月 26 日現在）.
3) SAS 公式サイト, http://www.sas.com/ .
4) R 公式サイト, http://www.r-project.org/ （2012 年 11 月 26 日現在）.
5) Nadeau, D. and Sekine, S.: A survey of named entity recognition and classification, Linguisticae Investigationes, vol. 30, no.1, pp. 3-26 (2007).
6) 金山 博, 那須川哲哉, 渡辺日出雄: 木構造変換を利用した評判分析手法, 人工知能学会論文誌 vol. 26, no. 1, pp. 273-283 (2011).
7) Twitter, https://twitter.com/ .
8) Twitter, サーチ API, https://dev.twitter.com/docs/using-search （2012 年 11 月 26 日現在）.
9) Twitter, ストリーミング API, https://dev.twitter.com/docs/streaming-apis （2012 年 11 月 26 日現在）.
10) IBM, InfoShphere BigInsighs 製品サイト, http://www-01.ibm.com/software/data/infosphere/biginsights/ （2012 年 11 月 26 日現在）.
11) IBM, IBM Content Analytics 製品サイト, http://www-06.ibm.com/software/jp/data/search/textmining.html （2012 年 11 月 26 日現在）.
12) Jaql 公式サイト, http://code.google.com/p/jaql/ （2012 年 11 月 26 日現在）.
13) Hastie, T., Tibshirani, R. and Friedman, J.: The Elements of Statistical Learning, Springer (2001).
14) Kanayama, H. and Nasukawa, T.: Unsupervised Lexicon Induction for Clause-level Detection of Evaluations, Journal of Natural Language Engineering, vol. 18, no. 1, pp. 83-107 (2011).
15) 羽室行信，岡田克彦: テキストマイニングを用いた株式銘柄センチメントの測定とポートフォリオの構築 ～ マーケット・ニュートラルアプローチ ～, 電子情報通信学会第 1 回テキストマイニングシンポジウム(2011).

古関　聰（正会員）
E-mail: akoseki@jp.ibm.com
1998 年，早稲田大学大学院理工学研究科電気工学専攻博士課程修了．同年，日本アイ・ビー・エム株式会社入社．東京基礎研究所に入所以降，Java just-in-time コンパイラ，J2EE アプリケーション，並列分散処理システム等の研究に従事．2004 年情報処理学会論文賞．博士（工学）．

金山　博（非会員）
E-mail: hkana@jp.ibm.com
2000 年, 東京大学大学院理学系研究科修士課程修了．同年，日本アイ・ビー・エム株式会社入社．東京基礎研究所に入所以降，構文解析，意味解析などの自然言語処理の基礎技術，および翻訳，評判分析，文書校正，質問応答などへの応用に関する研究に従事している．博士（情報理工学）．

坪井　祐太（正会員）
E-mail: yutat@jp.ibm.com
2002 年，奈良先端科学技術大学院大学前期課程修了. 同年日本アイ・ビー・エム株式会社入社. 2009 年奈良先端科学技術大学院大学後期課程修了. 博士（工学）．東京基礎研究所にて，テキストマイニングの研究開発に従事．2010 年情報処理学会論文賞，人工知能学会現場イノベーション賞受賞．言語処理学会会員.

平出　涼（非会員）
E-mail: rhirade@jp.ibm.com
2002 年，東京工業大学大学院総合理工学研究科物理情報システム創造専攻修士課程修了．同年，日本アイ・ビー・エム株式会社入社．東京基礎研究所に入所後，パフォーマンス解析，データ解析などの数理モデル解析に関する研究に従事している．

千葉　立寛（正会員）
E-mail: chiba@jp.ibm.com
2011 年，東京工業大学大学院・情報理工学研究科・数理計算科学専攻博士課程修了．同年，日本アイ・ビー・エム株式会社入社．東京基礎研究所にて，大規模データ向けの基盤ソフトウェア・並列分散システムに関する研究に従事．博士（理学）．

米持　幸寿（非会員）
E-mail: pandrbox@jp.ibm.com
1987 年，IBM 入社．メインフレーム・ソフトウェアの障害対応，ソフトウェア・エバンジェリストなど経て，スマーター・シティー関連ソリューション・デザイン．主にビッグデータやテキスト分析テクノロジーを担当．

野村　尚（非会員）
E-mail: nomurah0@jp.ibm.com
2002 年，英国 Cass Business School にて MBA 取得．2005 年，国内 SI 企業を経て IBM 入社．国内外の金融機関に対するコンサルティングおよびソリューション企画・開発に従事．主にビッグデータやテキスト分析ソリューションを担当．

投稿受付：2012 年 10 月 01 日
採録決定：2012 年 12 月 03 日
編集担当：守安　隆（東芝ソリューション）