経営学におけるテキストマイニングの可能性

――仮説構築志向の利用方法――

高　木　修　一・竹　岡　志　朗

富山大学紀要. 富大経済論集　　第64巻第2号抜刷（2018年12月）

富山大学経済学部

経営学におけるテキストマイニングの可能性
――仮説構築志向の利用方法――

高木　修一・竹岡　志朗

キーワード：テキストマイニング，経営学，機械学習

目次

１．はじめに

本稿の目的は，経営学研究におけるテキストマイニングの利用方法について検討することである。そのため，先行研究をレビューしたのち，仮説構築のためにテキストマイニングを利用する手法として，「概念総当たり検討法」を提案する。

2000 年代以降，コンピューターの性能向上や自然言語処理技術の発展に伴って，所謂テキストマイニングと呼ばれる手法を用いた研究が数多く行われるようになっている。図 1 は，「Cinii」および「Web of Science Core Collection」を用い，「テキストマイニング」および「text mining」をキーワードに論文数の検索を行った結果である。このグラフからわかるように，日本においては 2013 年を頭打ちに論文数の増加は停滞しているようであるが，海外においてはいまなお増加傾向にある[1]。

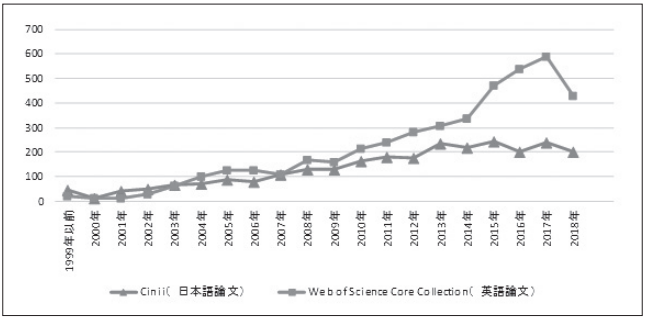

図１　テキストマイニングに関連する研究数の推移

日本語論文の増加傾向自体は収まりつつあるが，それは決してテキストマイニングへの興味が失われたからというわけではない。むしろ，様々な分野の雑誌において特集が編纂され，結果としてテキストマイニングという手法が普及したからであると考えられる。『人工知能学会誌』では 2001 年の 16 巻２号にて「テキストマイニング」というそのものズバリな特集が組まれている。『データベース白書 2005』には，「テキストマイニングの最新動向」という形で１章分の記述がある。『看護研究』は，2013 年の 46 巻５号にて「看護研究におけるテキストマイニング」という特集を組んでいる。情報経営の分野では，『日本情報経営学会誌』が, 2014 年の 35 巻１号にて「情報経営への言語的アプローチ」というタイトルで, テキストマイニングに関連した特集を行っている。『社会学評論』は，2017 年の 63 巻３号にて，「テキストマイニングをめぐる方法論とメタ方法論」という特集を編纂している。他にも，データマイニングの特集やビッグデータの特集でテキストマイニングが言及されていることもあり，非常に幅広い分野にてテキストマイニングが注目され，利用されていることがわかる。

先述したように，本稿の目的はこのようなテキストマイニングに対し，新たな利用法を提案するものである。より正確には，経営学研究において役立つと考えられる方法を検討ならびに提案する。本稿の方法が新規性を持つものであ

ると示すためには，既存のテキストマイニングを利用した経営学の研究がどのような特徴を持っているのか，どのような観点でテキストマイニングを利用しているのかを知る必要がある。

このような考えのもと，以降は次のような順序で議論を行う。まず，第２節では，経営学においてテキストマイニングを利用した先行研究のレビューを行う。そのレビューを踏まえ，テキストマイニングの利用法の特徴を明らかにする。続く第３節では，仮説構築を志向したテキストマイニング利用法を提案する。提案に際しては，実際にデータを用いた分析例も提示する。第４節では本稿の貢献と限界について述べる。

２．経営学におけるテキストマイニングの利用

本節では，経営学においてテキストマイニングを利用した研究のレビューを行う。レビューに際しては，研究に利用されたテキストの性質の違いという観点から，「テキストの収集目的とテキストマイニングの利用目的が一致するもの」,「テキストの収集目的とテキストマイニングの利用目的が一致しないもの」の２つに分けて整理する。

２．１　テキスト収集目的とテキストマイニングの利用目的が一致するもの

テキストの収集目的とテキストマイニングの利用目的が一致するものとは，コールセンターでの応答や自由記述アンケートなど，限定的な目的のために収集されたテキストを対象とする研究のことである。分析を行うことをある程度念頭においたテキストを対象とした研究ということもできる。

経営課題解決へのテキストマイニング利用を検討した，日本における初期の研究の１つとして位置づけられる那須川（2001）は，コールセンターにおける顧客と企業の対話として存在する問い合わせ記録を対象として分析することを試みている。この研究の特徴的な点は，コールセンターの問い合わせ記録（16,000 件），新聞記事（42,000 件），特許データ（2,000 件）を比較検討するといっ

た分析を行ったうえで，コールセンターにおける問い合わせの性質について議論を行っている点にある。那須川によれば，「コールセンターデータの分析は，さまざまな意味でテキストマイニングの理想的なアプリケーションである。第一に，データ自体が有益な情報を含んでいる。第二に，人手では手に負えない膨大なデータの分析という，従来できなかったことを可能とするという点で有用性が大きい。第三に，全体として傾向が重要なため，概念抽出における自然言語処理の多少の解析エラーはノイズとして無視でき，現在の自然言語処理技術のレベルで十分に効果をあげられる。第四に，分野が限定されているため意味辞書の構築が比較的低い労力で可能であり，適用開始時の負担が比較的少ない。（那須川，2001，p.225）」とされる。このように，コールセンターの問い合わせ記録の研究は，消費者と企業の対話というところに有益なデータがある，ひいては限定的な目的のためにとられたデータの中に有益な何かが眠っており，それをテキストマイニングによって発見できるという可能性を明らかにしたものだと捉えることができる。

その流れを組むものとして，マーケティング分野における研究，より正確には消費者の意見や視点を学術研究として明らかにすることを試みるものが多数出現することとなる。例えば，磯島（2006）は，「米の品質と価格に関する消費者意識調査」にある自由記述回答（452 件）を分析対象として，米に対する消費者意識の解明を試みている。ここでは，回答者属性別の出現語の集計や等質性分析（対応分析）などを行っている。石川・星野（2004）は，岡山県和気郡吉永町にある「八塔寺山荘」という町営宿泊施設にある落書き帳（146 件）のデータを「スプリング埋込み」と呼ばれる概念間の関係を可視化する手法を用いて，書き込みの内容と記入者の年齢との関係を分析，落書き帳の有効活用のための議論を行っている。これらの研究の特徴は，人力では困難あるいは人力では主観が入ってしまう自由記述に対し，テキストマイニングという手法を通すことによって研究者に依存する結果のぶれをなくそうとしている点にある。石川・星野（2004）は，「テキストデータの分析手法としては KJ 法が広

く知られている。複数のデータから新しい仮説の発見や枠組みの形成のための，カードを用いたデータ整理法である。しかしKJ法では，人がテキストを読んで内容を理解し，頭の中でマッチングを行い分類しなければならない。一度に処理できるカードの数には限界があり，膨大な量のテキストをカードとして処理することには不向きである。また，KJ法は分析者の主観と熟練度に依存するため，分析者による「ばらつき」が避けられない。（石川･星野, 2004, p.181）」と言及している。ここでは，定性研究で特に問題として議論されやすい，研究者の主観による分析結果のぶれという問題解決の方法としてテキストマイニングが用いられているという特徴がある。

２．２　テキスト収集目的とテキストマイニングの利用目的が一致しないもの

テキストの収集目的とテキストマイニングの利用目的が一致しないものとは，有価証券報告書やオンラインレビューのように，利用目的が事前に限定されていないテキストを対象として行う研究である。テキスト分析を行うことは研究者が独自に設定したに過ぎず，それ故にそれぞれの研究目的に応じて加工し，利用している研究であるとも言うことができる。

企業に近い情報という観点から言うならば，有価証券報告書のような定期的に発行されるテキストを利用した研究が挙げられるだろう。喜田(2006)は,「アサヒ」と「キリン」の22年分の有価証券報告書の「営業の状況」をテキストマイニングすることで，出現する概念数（名詞数）の変化と経営成果に関する変数（シェア，売上高，経常利益）との関係を分析している。その結果について，認知的組織革新研究の文脈から,「組織革新は組織的知識構造の変化を必要とする（喜田，2006，p.90）」という中心仮説並びに，複数の発見事実の関係の解釈を行っている。同じく，有価証券報告書を用いた分析として白田・坂上（2008）がある。白田・坂上は，倒産企業（20社）と継続企業（24社）の有価証券報告書に現われる語句をTF-IDF法を用いたテキストマイニングと，財務上の数値データを合わせて分析した。その結果，倒産企業と継続企業の間

には出現する語句の頻度に違いがあることなどがあきらかとなっている。海外においても Balakrishnan et al. (2010) は，1997 年から 2002 年の製造企業を対象とし 1,236 社の年次報告書（4,755 件）から業績を予想することを試みている。少し毛色は異なるが，Kleinman et al.(2017) は，７社の CSR レポートを対象に形式概念分析を用いることで，企業が持続可能性のための政策をどの程度順守しているのかを見極めることができるという議論を行っている。

企業の中に存在する従業員から得られたデータを用いた研究も存在する。安田・鳥山（2007）は，コンサルティング企業で用いられている業務用電子メールログ（約 37,000 通）を対象として分析を行った。この際，法的な問題，倫理的な問題を検討し，プライバシーの問題へ技術的な対処を行った上で研究を行っていると明言しているという特徴がある。なお，これらに配慮しつつ研究を行い，メールのやりとりの特徴的な用語，職位階層による違いやパフォーマンスによる違いなど，いくつかの属性情報を考慮することにより様々な結果を導いている。Speer(2018) は，米国の大規模金融機関の 2011 年から 2015 年までの間における，フルタイム従業員（約 5,000 人）の業績評価や離職率のような数値情報，そして直属の上司によるナラティブな評価（テキスト）を対象に分析を行っている。テキストに関しては，ワードクラウドによる可視化や Elastic Net 回帰を用いたスコアリングを行い，退職や昇格，昇給とどのように相関しているのかを分析している。その結果は多岐にわたるが，これまで従業員の業績評価においてあまり活用されてこなかった上司によるナラティブな評価をテキストマイニングによって自動的に分析することの価値が示されている。なおこの研究においても，匿名性を担保するために，すべての評定者（直属の上司）および非評定者（労働者）の氏名はランダムな ID に変更されていることも注記されている。

企業からは少し離れるが，オンラインレビューやオンラインコミュニティでのやり取りなど，インターネット上で生成されるテキストを対象とした研究も存在する。竹岡他（2014）は，インターネット掲示板「価格 .com」に記載さ

れたクチコミ（803,967 件）のデータを対象にイノベーションの普及を可視化するという観点から分析し，消費者が製品を認識する際には比較対象として選ぶ「ベンチマーク機種」や比較の際に特に利用する「優先概念」というものの存在を指摘している。この中で，データの性質について「予備調査を行った際，クチコミ件数が 500 件程度の場合は，１件のクチコミの持つ影響が大きく，分析結果に大きな誤差が生まれる可能性が高いことが分かった。この問題に対処するため，本稿ではクチコミ件数が 2000 件以上の機種に調査の対象を限定している。（竹岡他，2014，p.84）」という言及がある。一般的にテキストマイニングが平均的な傾向を明らかにする分析である以上，データ数が少数であることは結果にぶれをもたらすことがあるという指摘は，裏を返せば少数データではテキストマイニングといえども分析の精度に問題があると考えることの必要性を意味するだろう。

新製品開発に近づけた研究として，Christensen et al.(2016) は，オンラインコミュニティの 3,000 件のデータから，有益なアイデアを機械学習（教師付き学習，サポートベクターマシン）によって検出する方法について研究を行っている。視点は少し異なるが，Wei et al.(2010) は，DVD プレイヤーに関するオンラインレビュー（3,944 件）を用いて，形容詞を正と負の意見語として利用する製品特徴抽出の方法提案を行っている。新製品開発を行うプロセスにテキストマイニングを組み込むという視点でいえば，Shu et al.(2016) が，テキストマイニングを利用し，ユーザー主導での品質機能展開や頻度パターンツリーアルゴリズム，製品ロードマップを組み合わせた体系的な製品機能要件の優先順位付けの手法を提案している。

インターネット上のテキストのみではなく，テキスト以外の情報と組み合わせながら研究を行うことも行われている。三川他（2007）は，一般消費者へのインタビュー調査（22 名）と，「価格 .com」や「楽天市場」のユーザーレビュー（300 件）を用い，形態素解析を用いたうえで，顧客ロイヤルティの構造を明らかにしている。Chern et al.(2015) は，オンラインレビューが製品の販売数

にどのような影響を与えるのかをオンラインレビューの分類や線形回帰などからなる「eWord-of-Mouth Sales Forecasting Algorithm (WOMSFA)」というモデルを構築し，検証している。データとして台湾のベストセラー商品（107製品）に関する「UrCosme.com」のオンラインレビュー 100 製品（8,386 件のレビュー）を用いて検証している。この結果，クチコミが消費者行動（購買）に影響があることが検証されたが，一方でライフサイクルの長い製品ではほとんどオンラインでディスカッションが発生しないため，モデルの適合ができないという指摘も行っている。他にも，Lash and Zhao(2016) は，出演俳優の情報や映画の内容のテキストマイニングも含めた様々な情報を用いながら，映画の収益性の主要因を分析している。

インターネットにあるテキストから企業に有益な情報を引き出すというものとは違い，インターネットのテキストが生成される環境そのものへの貢献を志向する研究もある。Courssement et al.(2017) は，イノベーションコミュニティを対象として研究（10 個のイノベーションコミュニティ，1,611 名の投稿者，39,387 件の投稿）を行い，そこで投稿される文章と参加者の質や量に与える影響を分析している。Goes et al.(2014) は，オンラインレビューがどのように生成されるのかに着目し，ユーザーレビューの生成に何が影響しているのかを明らかにしている。Tussyadiah and Park(2018) は，「Inside Airbnb.com」に掲載されたデータセットを用いて，31,119 件の Airbnb のホストの記述情報を分析している。主に共起分析やクラスター分析などの分析を用い，ホストがどのような言葉を用いているのか，自己紹介のパターンにはどのようなものがあるのか，さらには自己紹介のパターンに対して旅行者がどのように反応するのかを議論している。Joorabch et al.(2016) は，StackOverflow に投稿された 186,000 件の Q & A 投稿を分析することによって，最も頻繁に尋ねられるトピックやテーマ（カテゴリ）を発見するための方法（有向グラフによる可視化や頻度分析）を提案した。これにより，彼らは最も学習者が直面しやすい困難を可視化することができたとする。

これらの他にも，より広く社会に存在する特許や論文を対象とする研究もある。酒井他（2009）は，特許明細書から技術課題情報を自動的に抽出するための手法を提案している。この研究の中では，358,085件の公開特許情報を利用し，精度や再現率などの点で検討を加えている。Lee et al.(2008) 特許情報に対し共起分析をベースとした様々な二次元情報化（マップ作製）を通じて，製品・技術ロードマップを作製するアプローチの提案を行っている。

Snehvrat et al.(2017) は，戦略経営論や組織論における両義性研究に関するレビューをテキストマイニングによって行っている。1997年から2016年の学術誌に掲載された論文の抄録（504件）をテキストマイニング（共起分析）することによって，どのような分野で研究が行われているか，将来的に注目を集めそうな分野はどこかなどを明らかにしている。White Ⅲ et al.(2016) は，国際戦略経営の研究分野において，736の論文を対象として文献分析を行い，研究動向の変化について明らかにしている。

２．３　経営学におけるテキストマイニングの利用の特徴と課題

これまで，テキストマイニングを用いた個々の研究を見てきた。ただし，当然ながらテキストマイニングを用いた研究の，レビュー研究自体も存在する。ここでは，それらレビュー研究の知見も活用しながら，経営学研究におけるテキストマイニングの利用の特徴について検討する。

日本におけるテキストマイニングについて，経営学以外も含めてレビューした上でアカデミックな方法論という観点から議論したものとして，喜田（2018）がある。喜田は，経営学のみならず，心理学や経済学などでのテキストマイニングを用いた研究をレビューした上で，「このように見てみると，マーケティング領域でのコールセンターにおける顧客の声分析から始まり，Twitterや口コミなど研究対象の広がりがある。つまり，ここでの動向から，①テーマの広がり，②研究対象の広がり，③分析手法の広がり，の３点が見られる。①は，様々なテーマで用いられるようになったことである。②は，口コミなどのインター

ネット上のデータを対象とするようになったことである。③はコレスポンデンス分析からニューラルネットワークや自己組織マップなどの機械学習の手法を用いるようになったことである。(喜田，2018，p.28)」という結論を導いている。完全に一致するものではないが，海外においても研究対象や手法の広がりは指摘されている。Usai et al.(2018) は，テキストマイニングによる知識発見に関する 85 の学術研究論文を対象にシステマティックレビューを行い，傾向として 1998 年から 2009 年の期間と 2010 年から 2017 年の期間の二つに大きく分けることができるとしている。前半と後半の大きな違いとしては，後半ではデータや手法の広がり（ユーザーレビューやマイクロブログデータ，センチメント分析など），対象分野の拡大（ビジネス，ファイナンスなど）があるとされている。

ただし，手法や対象は確かに広がっているものの，全体的な傾向と判断するべきなのかという点では少し疑問もある。斎藤（2011）は，喜田と同じく経営学も含めた多様な分野の先行研究をレビューした。その上で，分析手法に着目して整理を行い，「１．単語の出現頻度の集計，２．係り受けの頻度の集計，３．SOM，MDS による単語のマッピング，４．ベイジアンネット等による単語のネットワーク分析，５．コレスポンデンス分析，数量化Ⅲ類による単語と属性，対象の同時布置，６．対象のクラスタリング，７．SVM 等を用いたテキストの分類，８．キーワードの自動的な抽出」という形で分類を行っている。それに加えて，「確かにテキストマイニングには多くの応用事例がある。しかし，分析という観点から見ると，用いられている手法はかなり絞られている。単純な集計を行うか，テキストの属性の特徴について出現単語を用いて分析するといった，記述的な手法が大半を占めている。推測的な手法としては，機械学習によるテキストの分類が主となっている。(斎藤，2011)」とし，手法の限定性を指摘する。このような研究の偏りは，同じく斎藤が「個々の目的に合わせたデータの作り方をするのは応用研究を行うものには困難が大きい。ソフトウェアが示す手順通りに分析を行うとなると，基本的な名詞を抽出して単語間の関

連を見る，単語と属性の関係を見るという段階に留めざるを得ないのが現状なのだろうと思われる。また，一般的なデータマイニングと同様，統計的仮説検定等の基本的に推測統計の手法を用いることが難しい点も，応用事例が限られている理由であろう。(斎藤，2011)」と指摘する通りだろう。

また，データの作り方やソフトウェア開発の困難性という問題以外にも，テキストマイニングは根本的に大きな問題をいくつか抱えている手法である。例えば，ノイズに関しては，「テキストマイニングでは，前述のとおり，自然言語の曖昧性などからなるノイズを考慮する必要性が高い。そのため，マイニング結果を確認する際に，前処理の結果がどのような原文から得られたものか，その文脈はどうなっているのかを常に容易に確認できるようにしておくことが重要である。情報抽出結果を原文から切り離してしまうと，それが誤りであるかどうか確認することができず，信頼性の高い分析ができなくなってしまう。逆に言えば，常にインタラクティブに原文が参照でき，前処理の誤りなどが容易に特定できるようになっていれば，誤った判断を避けることができる。(人工知能学大事典，2017，p.682)」というような指摘がある。これは，先述した那須川（2001）や竹岡他（2014）における言及ともつながるものである。

かといって，ノイズが少ないであろうテキストとなれば，データへのアクセスの問題やテキストの量，法的・倫理的問題も発生する可能性が高い。安田・鳥山（2007）や Speer(2018) のように，企業内部から採取されたデータを用いる場合，法的・倫理的な問題について検討及び対処することが必要不可欠となる。テキストは世の中に数多く存在するが，その中身によっては，非常にセンシティブなデータになることに気を付ける必要がある。

このような課題を内包するテキストマイニングであるが，一方で情報の再利用や二次分析を行いやすいという利点もある。これまで先行研究をレビューしてきたことからわかる通り，テキスト収集の意図や目的はテキストマイニングによる研究をなんら制約しない。インターネット上にあるテキストから，書籍や新聞，報告書や会議資料などの灰色文献と呼ばれるものであっても分析が可

能である。

ここでレビューした先行研究のうちの多くは，多様なテキストを利用しながら，同時にテキスト以外を用いること，大規模データを用いること等で実証精度の担保をしていることは言うまでもない。しかし，果たして厳密な実証こそがテキストマイニングを利用するのに向いているのだろうか。テキストデータから得られた結果そのものを直接研究成果とすることだけがテキストマインイグの有効な利用方法なのだろうか。本稿の答えは否であり，その根拠として実証ではなく，仮説構築を志向するテキストマイニングの方法を以降で提案する。

3．仮説構築志向のテキストマイニングの方法

3．1　技術的背景

本稿で提案する仮説構築志向のテキストマイニングの方法である「概念総当たり検討法」であるが，過去の多くのテキストマイニングで用いられていた技術とは異なる手法を用いている。過去のテキストマイニングで用いられた手法の多くは，出現する単語の出現数を集計し，共起分析や分類を行うというというものである。サポートベクターマシンのような機械学習を用いた分析においても，その基礎としてあるのは単語の出現数であることが多い。このような単語数の集計による分析は，基本的に one-hot ベクトル表現のようなものであり，出現単語数に合わせて計算量が幾何級数的に増大する傾向にある。もちろん，コーディングルールによる情報の縮約などを行うことも可能であるが，結果的に研究者の主観の介在や手間暇を発生させることとなる。計算量の問題から実際には試行錯誤に多大な労力が必要となるため大規模なデータでは仮説構築を志向することが難しく，データを制約すれば単なる仮説検証あるいは実証になってしまうという問題が技術的に課された制約でもあった。

このような計算量の制約を解決する技術として，単語を数百次元のベクトルとして表現する分散表現によるテキストマイニングが近年利用されつつある。

これは，ある単語の意味に関して，前後数語ごとに切り分け，ニューラルネットによって次元の圧縮を行う方法である。分散表現を用いたテキストマイニングでは，同じ文脈で登場する語，つまり当該概念の周囲数語が同じ語の場合には似たベクトルが，同じ文章中で登場する語には全く異なるベクトルが与えられる。このように与えられた各単語のベクトルと単語間のベクトルのコサイン類似度を測定することで，分析対象となるテキストにおける出現語の意味の類似度を測定することができる。また，ベクトルの類似度の高い語間には交換可能性があり，単語を入れ替えても意味の通る文章が再現される可能性が高い。この方法によって，単語と単語の関係性をベクトル空間上に描写することができ，計算によって単語間の関係を明らかにすることができる[2]。

３．２　概念総当たり検討法

本稿では，「概念総当たり検討法」を提案する。なお，ここで用いるデータや手法自体はすでに竹岡(2018b)で公表したものである[3]。扱うデータは，「じゃらん net」に掲載されている東京ディズニーランド，東京ディズニーシー，ユニバーサル・スタジオ・ジャパン，横浜・八景島シーパラダイス，ナガシマスパーランドという５つのテーマパークに関するクチコミである。上記５つのテーマパークを分析の対象として選択した理由としては，データ収集時点で，これら５つがテーマパークカテゴリーの中で上位１位～５位であったことによる。データの収集は 2017 年 10 月 9 日から同 26 日にかけて行った。実際に学習に使用したデータは，各施設のクチコミ数に偏りがあり，このような不均衡データを用いた学習の結果には偏りが生じることが考えられることから，各施設のクチコミから 1,500 件をサンプリングした計 7,500 件である。

まず，単純に単語間の類似度をもとに５つの施設の類似度を計算すると，各施設の関係は下記表１のようになる。表１からは，「ディズニーランド」と「ディズニーシー」の類似度が高い，すなわちクチコミの中で類似する意味を持っていると考えられる。言い換えると，多数存在するクチコミ作成者の中では，同

じテーマパークであっても，「ディズニーランド」と「ディズニーシー」以外は比較的遠いものとして認識されていることが考えられる[4]。

表１　施設間の類似度

	ディズニーランド	ディズニーシー	USJ	横浜・八景島シーパラダイス	ナガシマスパーランド
ディズニーランド		0.72463	0.31852	0.17208	0.29442
ディズニーシー			0.26664	0.22879	0.24817
USJ				0.20728	0.25325
横浜・八景島シーパラダイス					0.25377
ナガシマスパーランド					

当然ながら，クチコミの中には施設名以外に無数の単語が含まれている。本稿で提案する「概念総当たり検討法」は，この無数に含まれる単語を大量に投入し，類似度の高い単語を抽出する。今回は，出現回数上位 200 位までの単語を総当たりで計算し，各施設と類似度の高い単語を抽出した。その結果が，表２である。この結果から，「USJ」と「大阪」，「横浜・八景島シーパラダイス」と「水族館」，「ナガシマスパーランド」と「アウトレットモール」など，単語を入れ替えても意味が通りそうなものが上位にきていることがわかる。すなわち，これらの単語が利用者の中では代替可能な言葉，類似する意味を持つ言葉として認識されているという仮説を考えることが可能となる。

この他にも，単語の加減算から仮説を検討しているのが，表３と表４である。表３は，各施設名からどのような単語を減算すると最も特徴が遠ざかるのかについて，出現回数上位 200 位までの単語で総当たりによって計算を行っている。「ディズニーシー」は「お酒」と「飲める」を減算すると，逆ベクトルの単語になり，「USJ」は「ハリーポッター」がなくなると，逆ベクトルの単語になっていることがわかる。すなわち，利用者にとって，「お酒が飲めること」こそが「ディズニーシー」の中核的な意味であり，「ハリーポッター」こそが「USJ」の中核的な意味であるという仮説が立てられるだろう。表４は，各施設に何を

加減算すれば，「ディズニーランド」に近似するのかを，出現回数上位 200 位までの単語で総当たりによって計算を行っている。その結果，「雰囲気」などの言葉が多いことから，利用者が認識するディズニーランドの最も特徴的な部分であり，他の施設と最も違う部分は，物的なものではなく「雰囲気」という非常に曖昧なものではないかという仮説を立てることができる。

表２　類似度による特徴抽出

ディズニーランド	
ディズニーシー	0.724639
雰囲気	0.612989
比べる	0.609007
飲める	0.588506
お酒	0.580588

ディズニーシー	
ディズニーランド	0.724639
お酒	0.651191
飲める	0.645726
雰囲気	0.530008
大人	0.514238

USJ	
年間パスポート	0.545035
Potter	0.513558
Harry	0.509493
大阪	0.5051448
行く	0.473982

横浜・八景島シーパラダイス	
水族館	0.641214
鎌倉	0.605503
シロイルカ	0.561197
金沢八景	0.519127
イルカ	516494

ナガシマスパーランド	
温泉	0.544516
ナガシマ	0.538653
地区	0.531
西	0.518665
アウトレットモール	0.513082

表３　減算による特徴抽出

ディズニーランド	
ディズニーシー , 飲める	-0.422241
ディズニーシー , お酒 ,	-0.412928
ディズニーシー , 雰囲気 ,	-0.395229
ディズニーシー , 違う ,	-0.36092
夢の国 , ディズニーシー	-0.335172

ディズニーシー	
お酒 , 飲める	-0.334084
お酒 , ディズニーランド	-0.326443
飲める , ディズニーランド	-0.319303
お酒 , クリスマス	-0.20991
飲める , クリスマス	-0.209498

横浜・八景島シーパラダイス	
水族館 , イルカ	0.0241484
水族館 , 海	0.0308357
イルカ , 海	0.0367161
ここ , イルカ	0.119278
ここ , 海	0.1281773

ナガシマスパーランド	
アウトレット , スチール	-0.019824
スチール , ドラゴン	-0.016169
アウトレット , ドラゴン	0.0190883
プール , スチール	0.045261
ジェットコースター , スチール	0.050169

USJ	
Potter, Harry	-0.168214
ハロウィン , Harry	-0.161839
Potter, ハロウィン	-0.160334
訪れる , Harry	-0.090942
Potter, 訪れる	-0.087906

表４　特徴の近似化

ディズニーシー	
ディズニーシー＋思う ＋雰囲気	0.7940271
ディズニーシー＋ 思う ＋ 違う	0.7886179
ディズニーシー ＋大人＋ ディズニー	0.7877395
ディズニーシー ＋ディズニー＋違う	0.7877291
ディズニーシー ＋ディズニー＋雰囲気	0.7876899

USJ	
USJ＋ 雰囲気＋ ディズニーシー	0.7415832
USJ＋ 大人＋ ディズニーシー	0.7340913
USJ＋ ディズニーシー＋ 飲める	0.7241877
USJ＋ 違う＋ ディズニーシー	0.7241855
USJ＋ 気＋ ディズニーシー	0.7176755

横浜・八景島シーパラダイス	
横浜・八景島シーパラダイス ＋ディズニーシー＋ 雰囲気	0.6634625
横浜・八景島シーパラダイス ＋ディズニーシー＋ 違う	0.6515939
横浜・八景島シーパラダイス ＋ディズニー＋ ディズニーシー	0.649789
横浜・八景島シーパラダイス ＋ディズニーシー＋ 飲める	0.6444271
横浜・八景島シーパラダイス ＋ディズニーシー＋ お酒	0.6360012

ナガシマスパーランド	
ナガシマスパーランド＋ディズニーシー ＋雰囲気	0.7373017
ナガシマスパーランド＋ディズニーシー ＋お酒	0.725242
ナガシマスパーランド＋ディズニーシー ＋飲める	0.7249891
ナガシマスパーランド＋ディズニーシー ＋違う	0.7165409
ナガシマスパーランド＋ 夢の国＋ ディズニーシー	0.7108018

本稿で示した分析結果は驚くべきものでは決して無いだろう。むしろ，当然の結果であり，この仮説自体には何ら面白みも新しさもない。重要なことは，このような仮説が得られる過程において，分析者の意図が介在せず，それでいて分析結果だけをみれば何らかの仮説を構築することができるというプロセス

である。紙幅の都合上，上位 200 単語すべてを表記していないが，実際には表２では５単語× 200 単語の関係（類似度）すべてが計算され，表４では５×８× 200^2 の計算結果が出現する。その結果からどのような仮説を構築するかはあくまで研究者の判断によるものだが，少なくとも仮説構築の基礎となるデータとして役立つのではないだろうか。

４．おわりに

本稿では，既存のテキストマイニング研究をレビューしたのち，仮説構築のためにテキストマイニングを利用する手法として，「概念総当たり検討法」を提案した。テキストマイニングはノイズや研究データへの接近という点で制約があるため，実証研究を志向すること自体に非常に高い壁がある。多くの研究ではその壁を乗り越えるために，テキスト以外のデータを使用することや，膨大な量のテキストを利用することなど行っているが，そのようなことができる場面は経営学研究では非常に限られたものである。それ故，本稿では実証ではなく仮説構築の道具としてテキストマイニングを利用することを考え，その手法として「概念総当たり検討法」を提案した。

今回示した「概念総当たり検討法」の例示は，１つのデータセットから計算した結果にすぎず，面白い仮説構築ができるということを十分に示したものにはなっていないだろう。しかし，先述したように，テキストマイニングの利点として情報の再利用や二次分析が行いやすいという特徴がある。今回の場合でも，旅雑誌や新聞広告の文章など，他のテキストを利用することで，より効率的な仮説構築ができる可能性がある。手作業で単語間の関係を見るのではなく，データと分析手法によって関係を表出化するという点が，本手法の最大の利点であろう。

最後になるが，本稿の貢献は，既存の研究をレビューした上で，テキストマイニングに内在する問題に立ち返り，その上で方法論としてテキストマイニングの別の可能性を示したことである。一方，限界としては，あくまで一手法を

示したに過ぎないことや，レビューが極めて限定的であることを挙げることができるだろう。まだまだ，テキストマイニングの技術自体が日進月歩で発展しているため，次の瞬間には別の可能性や方法が開かれる可能性はあるものの，現時点では仮説構築の道具としてテキストマイニングの利用に可能性があると考えられる。

謝辞

本研究は JSPS 科研費，若手研究（B）17K13787「イノベーションの普及過程で選考される意味属性のテキストマイニングによる可視化」の助成のもとで行っているテキストマイニングを用いた研究の一環としてなされたものです。

参考文献

石川修・星野敏（2004）「テキストマイニングを用いた都市農村交流ニーズの把握－岡山県吉永町ふるさと村の八塔寺山荘の落書き帳を対象として－」『農村計画学会誌』第23巻-suppl号，pp.181-186。

磯島昭代（2006）「テキストマイニングを用いた米に関する消費者アンケートの解析」『農業情報研究』第15巻，第1号，pp.49-60。

喜田昌樹（2006）「アサヒの組織革新の認知的研究－有価証券報告書のテキストマイニング－」『組織科学』第39巻，第4号，pp.79-92。

喜田昌樹（2018）『新テキストマイニング入門－経営研究での「非構造化データ」の扱い方－』白桃書房。

経済産業省商務情報政策局監修，財団法人データベース振興センター（2005）『データベース白書　2005』。

斎藤朗宏（2011）「日本におけるテキストマイニングの活用」『北九州市立大学ワーキングペーパーシリーズ』No. 2011-12。

酒井浩之・野中尋史・増山繁（2009）「特許明細書からの技術課題情報の抽出」『人工知能学会論文誌』第24巻，第6号，pp.531-540。

白田佳子・坂上学（2008）「人工知能アプローチによる「継続企業の前提」の解析－テキストマイニングによる非会計情報の分析－」（高田敏文編著『事業継続能力監査と倒産予測モデル』同文館出版，2008年）。

竹岡志朗 (2018a)「機械学習を活用したテキストマイニング - クチコミを用いた商品・サービスカテゴリーの横断分析 -」『桃山学院大学経済経営論集』第59巻，第4号，pp.101-122。

竹岡志朗（2018b）「機械学習を活用したテキストマイニング－特徴抽出の方法に関する検討―」『日本情報経営学会第76回全国大会予稿集』pp.155-158。
竹岡志朗・高木修一・井上祐輔（2014）「テキストマイニングを用いたイノベーションの普及分析」『日本情報経営学会誌』第35巻，第１号，pp.72-86。
那須川哲哉（2001）「コールセンターにおけるテキストマイニング」『人工知能学会誌』第16巻，第２号，pp.219-225。
三川健太・高橋勉・後藤正幸（2007）「テキストデータに基づく顧客ロイヤルティの構造分析手法に関する一考察」『日本経営工学会論文誌』第58巻第３号，pp.182-192。
安田雪・鳥山正博（2007）「電子メールログからの企業内コミュニケーション構造の抽出」『組織科学』第40巻，第３号，pp.18-32。
Balakrishnan, R., X. Y. Qui, P. Srinivasan (2010) "On the predictive ability of narrative disclosures in annual reports," *European Journal of Operational Research*, Vol.202, Issue 3, pp.789-801.
Chern, C. C., C. P. Wei, F. Y. Shen, Y. N. Fan(2015) "A sales forecasting model for consumer products based on the influence of online word-of-mouth," *Information Systems and e-Business Management*, Vol.13, Issue 3, pp.445-473.
Christensen, K., S. Norskov, L. Frederiksen, J. Scholderer (2016) "In Search of New Product Ideas: Identifying Ideas in Online Communities by Machine Learning and Text Mining," *Creativity and Innovation Management*, Vol.26, Issue 1, pp.17-30.
Goes, P. B., M. Lin, C. M. A. Yeung (2014) ""Popularity Effect" in User-Generated Content: Evidence from Online Product Reviews," *Information Systems Research*, Vol.25, No.2, pp.222-238.
Coussement, K., S. Debaere, T. D. Ruyck (2017) "Inferior Member Participation Identification in Innovation Communities: The Signaling Role of Linguistic Style Use," *Product Innovation Management*, Vol.34, Issue 5, pp.565-579.
Joorabchi, A., M. English, A. E. Mahdi (2016) "Text mining stackoverflow: An insight into challenges and subject-related difficulties faced by computer science learners", *Journal of Enterprise Information Management*, Vol. 29 Issue 2, pp.255-275.
Kleinman, G., C. H. Kuei, P. Lee (2017) "Using Formal Concept Analysis to Examine Water Disclosure in Corporate Social Responsibility Reports," *Corporate Social Responsibility and Environmental Management*, Vol.24, Issue 4, pp.341-356.
Lash, M. T. and K. Zhao (2016) "Early Predictions of Movie Success: The Who, What, and When of Profitability," *Journal of Management Information Systems*, Vol.33, Issue 3, pp.874-903.
Lee, S., S. Lee, H. Seol, Y. Park(2008) "Using patent information for designing new product and technology: keyword based technology roadmapping," *R&D Management*, Vol.38, Issue 2, pp.169-188.
Suh, Y., G. Kim, Seol, H. (2016) "Roadmapping for prioritisation of smartphone feature requirements based on user experiences," *Technology Analysis & Strategic Management*, Vol.29, Issue 8, pp.886-902.

Snehvrat, S., A. Kumar, R. Kumar, S. Dutta (2017) "The state of ambidexterity research: a data mining approach," *International Journal of Organizational Analysis*, Vol. 26 Issue: 2, pp.343-367.

Speer, A. B.(2018) "Quantifying with words: an investigation of the validity of narrative-derived performance scores," *Personnel Psychology*, Vol. 71, Issue 3, pp.299-333.

Tussyadiah, L. P., S. Park(2018) "When guests trust hosts for their words: Host description and trust in sharing economy," *Tourism Management*, Vol.67, pp.261-272.

Usai, A., M. Pironti, M. Mital and C. A. Mejri(2018) "Knowledge discovery out of text data: a systematic review via text mining," *Journal of Knowledge Management*, Vol. 22 Issue: 7, pp.1471-1488.

Wei, C. P., Y. M. Chen, C. S. Yan, C. C. Yang (2010) "Understanding what concerns consumers: a semantic approach to product feature extraction from consumer reviews," *Information Systems and e-Business Management*, Vol.8, Issue 2, pp.149-167.

White Ⅲ, G. O., O. Guldiken, T. A. Hemphill, W. He, M. S. Khoobdeh(2016) "Trends in International Strategic Management Research From 2000 to 2013: Text Mining and Bibliometric Analyses," *Management International Review*, Vol.56, Issue 1, pp.35-65.

提出年月日：2018 年 10 月 1 日

1　データ取得日は2018年９月17日である。そのため，2018年に関しては約９ヶ月分のデータであり，グラフ上は英語論文が大きく減少しているような形となっている。

2　本稿の目的からは外れるためこれ以上の説明は省略するが，詳細は竹岡（2018a）を参照して頂きたい。

3　正確には，データセットや単純な類似度計算については予稿にて文書化しているが，概念総当たり検討法については口頭発表に留まっており，文書としてまとめたのは本稿が初出である。

4　今回分散表現を算出するにあたってはfacebook researchの開発したfastTextを用いており，また学習手法としてはSkip-gramを使用している。fastTextのSkip-gramではsubword情報が活用されており，これによって未知語も分析の対象に含めることもできるが，その一方で字面の似た単語には近似するベクトルが算出されることになる。このため，ディズニーランドやディズニーシーのように字面がほぼ同じ単語に対しては過度に類似したベクトルが算出される可能性がある。