機械学習を活用したテキストマイニング（2）

―― 仮説の発見と検証 ――

竹　岡　志　朗

1．はじめに

近年，画像認識コンテストILSVRCにおける躍進や，AlphaGoの登場，そして様々な産業分野での応用事例の報告もあり，AIが社会的な関心を集め，機械学習やディープラーニングといったAIに関する用語も広く認知されるようになってきた。経営学研究および経営実践への応用可能性としても，近年の自然言語処理技術の向上，特に分散表現に関する技術が進歩したことで，十分に使用可能な技術となりつつある。

本研究では，このようなAI技術に含まれる機械学習技術を用いて商品・サービスの特徴をテキストマイニングによって分析・可視化する手法，特に仮説の発見と検証に関する方法を提案する。今回提案するテキストマイニングの手法は，現在主流の計量テキスト分析で用いられる単語等の集計値に基づくものではなく，機械学習によって算出される単語の分散表現に基づくものである。この方法を用いることによって，消費者の経験とその過程で構成される意味に基づいた分析が可能となる。本稿が提案する方法は経営学研究者にも有益だが，実務家にとっても新しい商品・サービスの企画やモデルチェンジ時に，他社の商品サービスとの比較がこれまで以上に容易になり，より詳細な分析をもとに実務を進めることができると考えられることから，有益だと考えている。分散表現に基づいたテキストマイニングは，まだまだ未完成の技術であり，定まった手法ではないが，今後もますます発展が期待される分野である。

キーワード：テキストマイニング，機械学習，分散表現，fastText，クチコミ

2．分散表現を用いたテキストマイニング

2．1　計量テキスト分析と分散表現テキストマイニング

現在のテキストマイニングの主流は「計量テキスト分析（樋口，2014）」と呼ばれるもので「計量的分析手法を用いてテキスト型データを整理または分析し，内容分析を行う方法（樋口，2014，p. 15）」である。その基本は文章を分かち書きし，出現する単語を集計すること，より進んだ分析としては文章内での単語の共起関係やJaccard係数を算出するなどし，その結果をもとに単語間の関係を共起ネットワークや階層的クラスターとして描きだし，考察することである。

他方，今回提案する手法は，自然言語処理の分野で発展した機械学習の技術を基礎にしたテキストマイニングであり，計量テキスト分析と区別するために，本稿では分散表現テキストマイニングと呼ぶものである。分散表現テキストマイニングは自然言語処理の分野で発展した分散表現に関する技術を応用したもので，図 1 のように，文章や単語を 100～300 次元程度の分散表現（ベクトル表現）に変換し分析を行う（図 1 は簡略化のため 3 次元で描写）。この技術を用いることで単語間の意味の類似度を単語の分散表現間のコサイン類似度として測ることが可能となり，これまでの計量テキスト分析では困難だった意味に基づくテキストマイニングが可能となる（竹岡，2018）。

自然言語処理の分野で単語の分散表現に関する研究が進んだ背景としては，機械翻訳，要約，検索などの技術の分野では，例えば「PC」と「パソコン」のように，表層的な単語の字面は違っていても，意味的に等価であれば同じものとして扱いたいという要求があり，これに対応するためには単語の類似度を正確に測定する必要があった。Bengio et al.（2003）でニューラル・ネットワークを用いて分散表現を算出する方法が提案される。しかし，この方法はデータ量が多ければ多いほど精度が向上するが，自然言語処理で通常用いられる程度の語彙数を学習させようとすると，計算数の激増により現実的なものではなくなるという問題があった。その後，Mikolov et al.

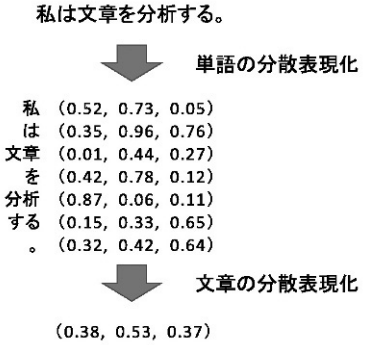

図1　単語と文章の分散表現化

(2013) で負例サンプリングを採用したWord 2 vec[1)]（自然言語処理ソフトウェア）が提案され，現在では Joulin et al. (2016) の fastText[2)]が多方面で使用されている。Word 2 vec や fastText で計算された分散表現を用いれば，単語の類似度を測るだけではなく，単語間の関係を加減算できること，さらには，「韓国とソウル」や「中国と北京」のように，対となる単語群の間にはオフセット関係があることも分かっている（図 2)。

このような単語をベクトル表現化する技術の背景にあるのが分布仮説 (Harris, 1954；Rubenstein & Goodenough, 1965）である。分布仮説とは「単語の意味はその単語が使われた際の周囲の単語によって決まる（鈴木他，2017)」という考えであり，Word 2 vec やfastTextはこの仮説を背景に単語の分散表現，つまりベクトルを算出している。例えば「晩　ごはん　は　焼き鳥　です」,「晩　ごはん　は　焼肉　です」,「朝　ごはん　は カレーパン です」の三文から各出現語のベクトルを算出する場合,「焼き鳥」と

1）「Word2vec」という用語はMikolov et al. (2013) の開発したソフトウェアを指す場合と，Skip-gramやCBOWなどの技術を用いた分散表現化を指す場合がある。本稿ではWord2vecをソフトウェア名として使用している。

2）fastTextはFacebook社が開発したオープンソフトウェアである。fastTextと同様に分散表現を算出するソフトウェアとしてはWord2vecやGloveなどもあるが，fastTextは他のソフトウェアと比較して正確性や速度の点で優れているとされている。

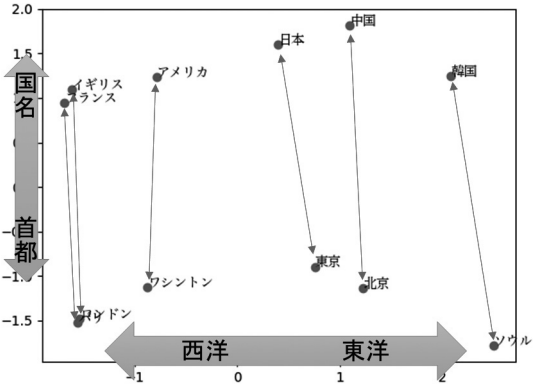

図 2　単語のオフセット関係[3)]

「焼肉」は同じ文脈語（「晩」と「ごはん」）から計算されるのでベクトルはよく似たものになり，他方で「カレーパン」の場合，文脈語が「朝」と「ごはん」なので「カレーパン」と「焼き鳥・焼肉」は，少し異なるベクトルが算出される。また，「焼き鳥」と「焼肉」のベクトルのコサイン類似度を計算した場合，その計算結果は「焼肉」と「カレーパン」のコサイン類似度と比較して高いものとなる（図 3）。

分散表現の算出では，より多くの文章を集めることができれば，類似する単語，単語の登場する文脈といったより多くの情報を学習に使用でき，結果として単語の類似度を正確に計算することができるようになる。

このような分散表現をテキストマイニングの基礎技術として使用することで，これまでには難しかった意味に基づいた分析が可能になる。計量テキスト分析の基本は，先述の通り，単語の出現回数や共起関係の強さの集計が基本であった。しかし，出現回数や共起関係の強さを調べただけではその単語に付与された意味の分析はできない，つまり「○○という単語が×回，△△という単語が×回と，同じ×回登場しているので○○と△△は同じ意味」と

3）Wikipediaで学習したモデルを使用。国名と首都名をピックアップし，PCAで座標を 300 次元から 2 次元に圧縮しプロットした。

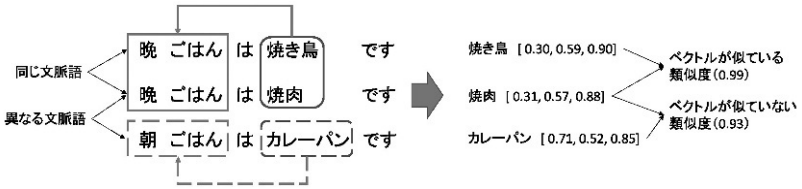

図3　単語の分散表現（ベクトル表現）化とコサイン類似度

はならないし，「○○という単語が△△と一緒に登場する回数，○○と□□が一緒に登場する回数が同じ×回なので，△△と□□が同じ意味」ともならない。しかし，分散表現を用いると類似度の高いベクトルを持つ単語は意味が似ていることが分かっているので，意味に基づいてテキストを分析できるようになる。

また，その分析に消費者のクチコミという体験に基づく言葉を学習データとして用いることで，これまでの外形的評価基準（敷地面積や入場料金のようなもの）とは異なる，消費者の心理的体験によって生まれた内的評価基準に基づいて用いられた言語表現を分析に使用できるという利点が生まれる。

2．2　代理変数を用いた特徴の抽出

しかし，類似度の測定や，単純な「類推問題（Mikolov et al, 2013）」を解く[4]だけではテキストマイニングには不十分である。そこで，本稿では分散表現をテキストマイニングの技術として応用するために，分析対象となる商品やサービスの特徴を，特徴となる語と商品・サービス名の類似度で代理する，つまり商品・サービスの持つ特徴の代理変数として類似度を用いる手法を採用する。具体的にはクチコミ中に登場する施設名と様々な出現語の類似度を施設の特徴の代理変数として使用することで，施設の特徴を可視化する手法を採用する。ここで特徴とは「近い」や「安い」，「デート」，「旅行」など文章中に出現する単語を指している。

4）高木・竹岡（2018）では，類推問題によるテキストマイニングによってテーマパークの特徴を可視化している。

たとえば，5 節で分析に使用する学習済みモデルを用いると，「海遊館」という単語と「デート」という単語の類似度は 0.45 で「旅行」との類似度は 0.30，「沖縄美ら海水族館」と「デート」の類似度は 0.23 で「旅行」との類似度は 0.41 である。この計算結果から，海遊館と沖縄美ら海水族館の相対的な特徴としては，沖縄美ら海水族館が海遊館と比較して旅行客が訪れる施設という特徴があり，他方で海遊館がデートで訪れられる施設という特徴があるという可能性を推測することができる（図 4）。このような分析は，単語が様々な成分から構成されており，それら成分と，その影響の程度によって意味が決まっていることによって可能となる。

以下では，分散表現テキストマイニングによって二つの分析を行う。4 節では，20 のレジャー施設に対するクチコミを対象に，すでにある仮説を検証するスタイルの分析を行う。5 節では，上記 20 施設の中から水族館カテゴリーに含まれる 5 施設に対するクチコミを用いて，仮説の発見から検証といったスタイルの分析を行う。

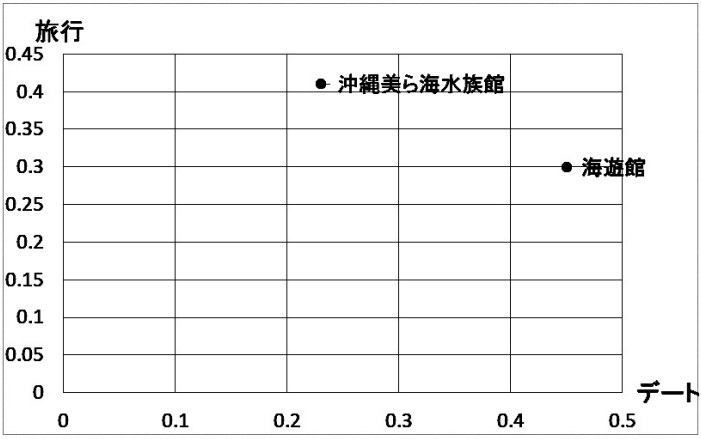

図 4　代理変数を用いた各施設の特徴の可視化

3．分析に使用するデータと方法

4節，5節ではインターネット上のクチコミデータを用いて分析を行っていく。使用したデータは「じゃらんnet」に投稿された「テーマパーク・レジャーランド（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）」，「水族館（沖縄美ら海水族館，鳥羽水族館，鴨川シーワールド，海遊館，名古屋港水族館）」，「動物園・植物園（アドベンチャーワールド，旭川市旭日山動物園，神戸市立王子動物園，東山動植物園，上野動物園）」「アウトレットモール（神戸三田プレミアム・アウトレット，三井アウトレットパーク　ジャズドリーム長島，三井アウトレットパーク　木更津，三井アウトレットパーク　滋賀竜王，三井アウトレットパーク　マリンピア神戸）」の4つのカテゴリーのランキング上位5施設，計20施設に関するクチコミ計71218件である。「じゃらんnet」に投稿されたクチコミ情報はインターネット上に投稿されている情報の一部でしかなく，本来であればインターネット上に投稿されているすべての情報を分析する必要があるが竹岡・高木（2018）に従い，今回はじゃらんnetの情報だけを分析の対象とした。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。

クチコミデータに関してはカタカナを全角に，アルファベットと数字を半角に変換，当該施設名に関わるものは，分析の中心的単語であることを考慮し表記ゆれの修正といった前処理を行った。分析には「Vector to[5]」を用いた。

実際に分析したデータは，各施設のクチコミ数に偏りがある。このような不均衡データを用いた学習では結果に偏りが生じることが考えられることから，4節では各施設のクチコミから1000件をサンプリングし，計20000件

5）Vector toは筆者が作成したfastTextをバックグラウンドで使用するフリーのテキストマイニングソフトウェアである。fastText単体では分散表現を用いた類似度の計算や単純な類推問題を解くことはできるが，テキストマイニングとしては少し物足りない点がある。Vector　toでは分散表現を様々な方法で活用することで，分散表現のテキストマイニングへの応用可能性を拡張している。（URL: https://vector-to.osdn.jp/）

のクチコミを，5 節では各施設から 2000 件をサンプリングし 10000 件のクチコミを分析に用いた[6]。

4．仮説の検証[7]

4 節では既にある仮説を，分散表現テキストマイニングを用いて検証する。まず，水族館プロデューサー中村元氏のテレビ番組での下記の発言[8]から仮説を構築する。

> 「水族館というのはデートに行くには一番いい場所なんですよ。（中略）たとえば，動物園だったりとか，テーマパークだったりすると，夏は，女性は汗だらだらで，化粧取れるから嫌だとかさ，あとハイヒール履けないから嫌だとか，オシャレできなかったりするじゃないですか。でも，水族館ってなんとなくちょっと知的で健全で涼しくて，オシャレもできるし，という感じで。（後略）」

この発言から検証する仮説として①水族館はテーマパークなどと比較してデートで訪れる場所である，その理由としては汗をかかずに楽しめる，つまり②エアコンが効いているためである，の 2 つを立てる。

表 1 は各施設名と「デート」の類似度を低いものから高いものへ並べたものである。水族館は斜字で記載し，先頭に★記号を付している。

これを見ると，水族館と「デート」の類似度が高い傾向にある，つまり水族館という施設のデート属性が高いことがわかる。次に表 2 は各施設名と「旅行」との類似度である。

「旅行」はテーマパークとの類似度は高いが，水族館とはそれほど高くな

6）分散表現化する際のハイパーパラメーターは次のとおりである（学習方法，次元数，学習回数，学習率，文字N-gramの順で記載）。4 節はsupervised，100，50，0.05，0，5 節はSkip-gram，100，30，0.05，2～4。
7）本節におけるデータ等は日本情報経営学会第 77 回全国大会での発表資料を再構成したものである。
8）株式会社TBS制作，2017 年 5 月 30 日放送の「マツコの知らない世界」より。

いことがわかる。「デート」と「旅行」の相関係数は 0.179 と低く，これら各施設間，あるいはカテゴリー間にはデート属性と旅行属性において違いがあることが推測される（図 5）。

表 1　各施設名と「デート」の類似度

施設名	デートとの類似度
三井アウトレットパーク木更津	−0.5497
三井アウトレットパーク滋賀竜王	−0.4260
ディズニーランド	−0.3163
★鴨川シーワールド	−0.2496
ディズニーシー	−0.2035
ナガシマスパーランド	−0.1644
旭川市旭山動物園	−0.1003
上野動物園	−0.0934
三井アウトレットパークジャズドリーム長島	−0.0216
神戸市立王子動物園	−0.0003
アドベンチャーワールド	0.0615
神戸三田プレミアム・アウトレット	0.0635
★沖縄美ら海水族館	0.0779
★鳥羽水族館	0.1102
三井アウトレットパークマリンピア神戸	0.1617
USJ	0.1924
東山動植物園	0.2372
横浜・八景島シーパラダイス	0.3909
★名古屋港水族館	0.4638
★海遊館	0.7003

表 2　各施設名と「旅行」の類似度

施設名	旅行との類似度
神戸三田プレミアム・アウトレット	−0.5327
三井アウトレットパークマリンピア神戸	−0.4731
★鴨川シーワールド	−0.3812
東山動植物園	−0.3056
★名古屋港水族館	−0.1993
三井アウトレットパーク滋賀竜王	−0.1738
ディズニーシー	−0.1658
ナガシマスパーランド	−0.1438
三井アウトレットパークジャズドリーム長島	−0.1253
三井アウトレットパーク木更津	−0.0904
神戸市立王子動物園	−0.0364
★海遊館	0.0766
旭川市旭山動物園	0.0887
上野動物園	0.1046
★鳥羽水族館	0.1704
★沖縄美ら海水族館	0.1844
USJ	0.2346
アドベンチャーワールド	0.3403
ディズニーランド	0.4515
横浜・八景島シーパラダイス	0.5264

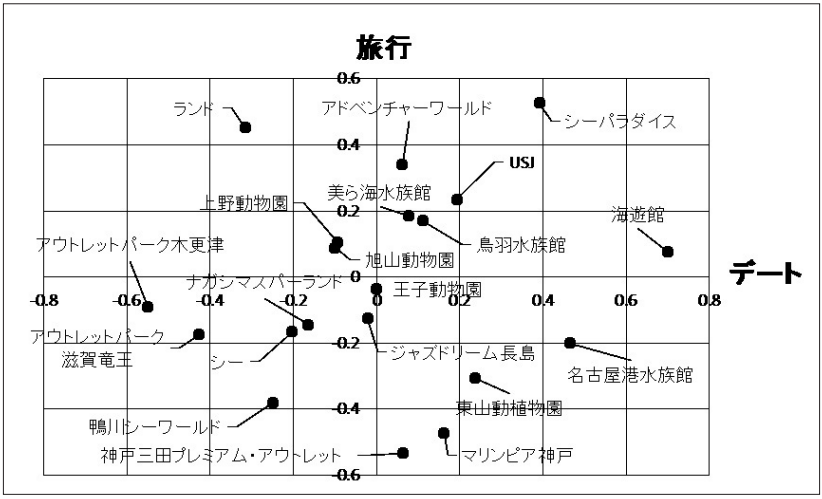

図 5　各施設における「デート」と「旅行」の知覚マップ

次に仮説の②である。夏場に汗をかかないのはエアコンが効いた施設内での行動が多いためと考えられる。表 3 は各施設名とエアコンの類似度である。水族館は全体として類似度が高い傾向にある。

図 6 は各施設名と「エアコン」・「デート」の類似度を知覚マップとしてプロットしたものである。「エアコン」と「デート」の相関係数は 0.650，p値 0.002 と相関関係にあることがわかる（同様に，「エアコン」と「旅行」の相関を計算したところ 0.144，また「エアコン」と「家族」では－0.092 と低いものであった）

以上消費者のクチコミを分析した結果から，仮説①および②はその妥当性が高い，つまり水族館が他の施設と比べてデートに向いていることが推測される。

5．仮説の発見と検証

本節は，水族館 5 施設に分析対象をしぼり，分散表現に加えて外形的データ（仕様など）を用いる，つまり単語の類似度と外形的データを併用するこ

表３　各施設名と「エアコン」の類似度

施設名	エアコンとの類似度
三井アウトレットパーク滋賀竜王	-0.588
神戸市立王子動物園	-0.411
ディズニーシー	-0.355
ナガシマスパーランド	-0.328
東山動植物園	-0.316
ディズニーランド	-0.256
三井アウトレットパークジャズドリーム長島	-0.241
三井アウトレットパーク木更津	-0.229
USJ	-0.090
旭川市旭山動物園	-0.083
上野動物園	-0.011
アドベンチャーワールド	0.104
★鴨川シーワールド	0.133
三井アウトレットパークマリンピア神戸	0.258
神戸三田プレミアム・アウトレット	0.321
横浜・八景島シーパラダイス	0.326
★沖縄美ら海水族館	0.329
★海遊館	0.469
★名古屋港水族館	0.771
★鳥羽水族館	0.779

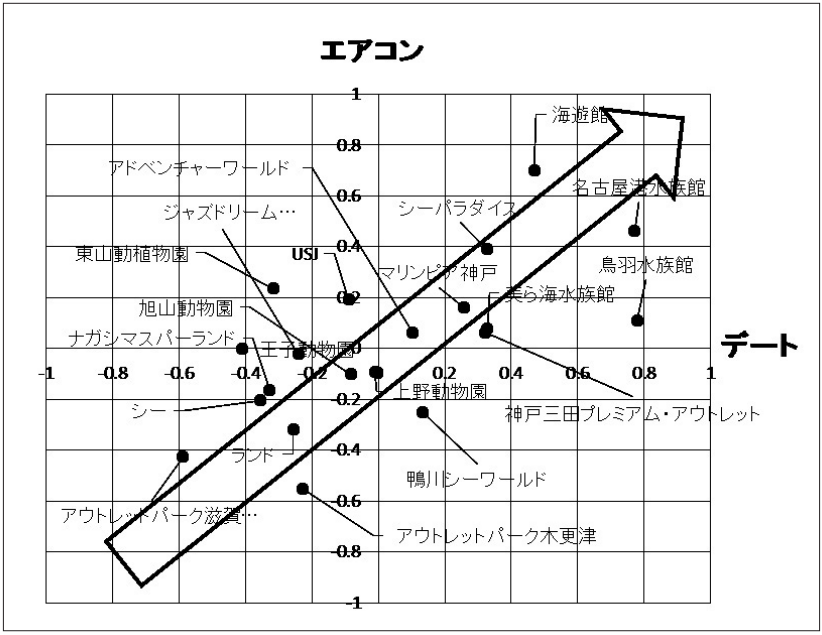

図６　各施設における「デート」と「エアコン」の知覚マップ

とで，消費者の体験によって構築される意味と，客観的仕様から各施設の特徴を可視化したうえで，仮説を発見・検証する手法を提案する。外形的データとしてはWikipedia[9]に掲載されている来場者数，飼育種類数，延べ床面積を利用した。

5．1　外形的データの併用による仮説の発見[10]

まず，外形的データを用いた仮説の探索を行う。下記では，水族館に関するクチコミ 10000 件の中で上位出現回数 200 位までの単語を抽出（表 4，5，6 中最左列，3 行目以下），それらと先述 5 施設名の類似度を計算する（同右側の 5 列，3 行目以下）。この計算された類似度と，各施設の外形的データ（同右側の 5 列，2 行目）の相関係数（同左 2 列目，3 行目以下）を総当たりで計算した。表 4〜6 は相関の強い 20 概念を抽出したものである[11]。

表 4 より，代理変数を用いて算出した各施設の特徴と各施設の延べ床面積から以下のことがいえる。

- 「旅行」と延べ床面積は逆相関の関係にある（旅行と延べ床面積の相関係数は−0. 993（以下同じ））
  - ➡　広い水族館は旅行で行くところではない

9）https://ja.wikipedia.org/wiki/日本の水族館（最終確認日：2018 年 9 月 28 日）。名古屋港水族館の飼育種類数についてはhttps://ja.wikipedia.org/wiki/名古屋港水族館（最終確認日：2018 年 9 月 28 日）

10）本項におけるデータ等は日本経営学会第 92 回大会での発表資料を再構成したものである。

11）表中の破線下はp値が 0. 05 を超えるものであり，本来は相関があるとは言えないものである。しかし，5 件と少ない中での相関係数の算出では，p値は必然的に大きいものとなり，相関係数は高くとも相関があるとは言えないものが多く登場する。本研究では統計的な厳密さは求めておらず，水族館というサービスカテゴリーを事例に分散表現テキストマイニングによって現状を可視化することの可能性を検討することを目的としているため，厳密には相関があるとは言い切れなくとも相関係数が高く算出されていることを考慮し，掲載することとした。
　このような問題は実務家が本研究と同様の方法で分析を行った際にも起こりうるものである。どの程度の統計的厳密性をもって分析を行うのかといった判断はその状況によって異なるだろうが，意思決定を支援するための分析という観点でこのような分析を行う場合には，統計的厳密性よりも，総当たりで様々な概念を分析できる利点を活かす方がよいと思われる。

表4　延べ床面積との相関

	相関	p値	分散	海遊館	沖縄美ら海水族館	鴨川シーワールド　鳥羽水族館	名古屋港水族館
延べ床面積（㎡）				31044	19199	22699	24981	41529
旅行	*−0.993*	0.001	0.008	0.298	0.414	0.366	0.371	0.180
楽しむ	*−0.984*	0.002	0.005	0.179	0.256	0.235	0.194	0.072
一緒	0.935	0.020	0.006	0.231	0.139	0.127	0.121	0.291
利用	−0.909	0.032	0.005	0.191	0.277	0.181	0.212	0.084
前	−0.905	0.035	0.003	0.184	0.204	0.196	0.162	0.062
有名	−0.904	0.035	0.002	0.330	0.397	0.347	0.328	0.284
最高	*−0.903*	0.035	0.013	0.148	0.294	0.368	0.246	0.079
くださる	−0.903	0.036	0.014	−0.022	0.231	0.183	0.157	−0.020
小さい	0.900	0.037	0.005	0.228	0.151	0.137	0.243	0.318
家族	*0.886*	0.045	0.004	0.367	0.236	0.328	0.298	0.395
お勧め	−0.845	0.072	0.012	0.017	0.274	0.217	0.159	0.052
過ごせる	0.843	0.073	0.010	0.234	0.030	0.015	0.114	0.215
良い	−0.829	0.083	0.005	0.269	0.262	0.328	0.266	0.141
以上	−0.817	0.091	0.005	0.258	0.247	0.240	0.270	0.105
凄い	−0.813	0.094	0.005	0.231	0.314	0.190	0.271	0.123
カップル	*0.796*	0.107	0.025	0.433	0.096	0.076	0.180	0.352
目	−0.774	0.124	0.008	0.160	0.288	0.109	0.158	0.042
言う	−0.742	0.151	0.004	0.299	0.347	0.411	0.287	0.252
デート	0.712	0.178	0.020	0.448	0.234	0.147	0.133	0.390
写真	−0.711	0.178	0.002	0.179	0.206	0.130	0.171	0.098

➡　あるいは，大都市内の水族館の延べ床面積が広い傾向にあるためか

- 「楽しむ」や「最高」と延べ床面積は逆相関の関係にある（楽しむ：−0.984，最高：−0.903）

  ➡　広いと楽しめないということか

- 「家族」や「カップル」と延べ床面積は相関関係にある（家族：0.886，カップル：0.796）

  ➡　大都市内の水族館なので家族やカップルには手軽ということか

続いて，表5では外形的データに飼育種類数を用いて相関関係を分析する。

- 「面白い」や「珍しい」は飼育種類数と相関関係にある（面白い：0.979，珍しい：0.884）

  ➡　飼育種類数が多いと必然的に珍しい動物を飼育することになり，それが来場者に面白いと認識され，相関関係が高くなっているか

- 「ジュゴン」や「セイウチ」が飼育種類数と相関関係にある（ジュゴン：0.953，セイウチ：0.895）

  ➡　日本では鳥羽水族館だけがジュゴンを飼育しているため，鳥羽水族

表 5　飼育種類数との相関

	相関	p値	分散	海遊館	沖縄美ら海水族館	鴨川シーワールド	鳥羽水族館	名古屋港水族館
飼育種類数(種)				580	740	800	1200	500
面白い	*0.979*	0.004	0.007	0.065	0.105	0.162	0.244	0.028
ジュゴン	*0.953*	0.012	0.033	0.231	0.236	0.262	0.627	0.171
暑い	−0.948	0.014	0.003	0.201	0.215	0.189	0.091	0.233
見学	0.924	0.025	0.002	0.078	0.140	0.153	0.189	0.096
笑	0.905	0.035	0.003	0.176	0.186	0.159	0.258	0.118
みる	0.899	0.038	0.006	0.177	0.293	0.228	0.368	0.206
すぎる	0.898	0.038	0.003	0.091	0.161	0.191	0.239	0.137
セイウチ	*0.895*	0.040	0.026	0.123	0.108	0.101	0.464	0.094
珍しい	*0.884*	0.046	0.011	0.173	0.188	0.171	0.416	0.186
場所	−0.884	0.047	0.002	0.277	0.240	0.193	0.162	0.244
飼育	0.883	0.047	0.011	0.076	0.131	0.205	0.354	0.165
動物	0.872	0.054	0.012	0.175	0.103	0.257	0.362	0.110
見応え	0.856	0.064	0.001	0.144	0.159	0.154	0.229	0.166
来る	0.838	0.076	0.010	0.293	0.406	0.274	0.480	0.244
ラッコ	0.838	0.076	0.014	0.173	0.091	0.184	0.399	0.150
混雑	−0.835	0.079	0.001	0.146	0.159	0.171	0.098	0.179
大変	0.833	0.080	0.006	0.110	0.078	0.167	0.200	−0.003
食事	0.830	0.082	0.001	0.157	0.124	0.174	0.204	0.119
たくさん	0.828	0.083	0.007	0.283	0.266	0.225	0.428	0.238
デート	−0.824	0.086	0.020	0.448	0.234	0.147	0.133	0.390

館の類似度が高くなっているか

➡ セイウチは鳥羽水族館と鴨川シーワールドで飼育されているが鳥羽水族館だけが類似度が高くなっており，これはショーを行っているかどうかの差があらわれているか

最後に表 6 では来場者数との相関を算出したものである。

- 「近い」と来場者数は逆相関の関係にある（近い：−0. 977）
  - ➡ この「近い」が家から「近い」ことを指しているのか，あるいは人と動物の距離が「近い」ことを指しているのかはわからない
- 「きれい」や「優雅」と来場者数は相関関係にある（きれい：0. 948，優雅：0. 944）
- 「巨大」や「規模」，「大きい」と来場者数が相関関係にある（巨大：0. 877，規模：0. 865，大きい：0. 859）
  - ➡ 延べ床面積と「最高」や「楽しむ」は逆相関の関係にあったことから，「巨大」，「規模」，「大きい」は水槽のサイズを指している可能性が高い（「水槽」と来場者数も相関関係にあるため）
- 来場者数は「夕方」と相関関係にある（夕方：0. 893）

表6　来場者数との相関

	相関	p値	分散	海遊館	沖縄美ら海水族館	鴨川シーワールド	鳥羽水族館	名古屋港水族館
来場者数（万人）				217	281	80	83	199
スポット	0.981	0.003	0.011	0.313	0.341	0.116	0.117	0.240
近い	*−0.977*	0.004	0.003	0.132	0.069	0.211	0.189	0.136
きれい	0.948	0.014	0.012	0.209	0.313	0.047	0.106	0.268
優雅	0.944	0.016	0.019	0.299	0.421	0.091	0.118	0.186
水槽	*0.932*	0.021	0.025	0.381	0.429	0.061	0.138	0.208
混む	0.914	0.030	0.005	0.240	0.257	0.130	0.084	0.157
サメ	0.913	0.030	0.025	0.307	0.277	−0.032	−0.006	0.108
何度	0.904	0.035	0.004	0.341	0.402	0.230	0.305	0.338
泳ぐ	0.903	0.036	0.016	0.296	0.396	0.127	0.103	0.158
アシカ	−0.896	0.039	0.021	−0.018	−0.002	0.221	0.321	0.150
ジンベイザメ	0.894	0.041	0.043	0.448	0.584	0.139	0.126	0.192
夕方	*0.893*	0.041	0.020	*0.286*	*0.307*	0.092	−0.035	0.136
眺める	0.892	0.042	0.014	0.132	0.266	−0.058	0.067	0.091
動物	−0.879	0.050	0.012	0.175	0.103	0.257	0.362	0.110
巨大	*0.877*	0.051	0.045	0.431	0.509	0.098	0.044	0.129
違う	0.871	0.055	0.002	0.292	0.366	0.246	0.261	0.270
規模	*0.865*	0.058	0.004	0.304	0.300	0.161	0.230	0.300
大きい	*0.859*	0.062	0.008	0.407	0.422	0.192	0.312	0.329
施設	0.855	0.065	0.005	0.295	0.307	0.138	0.205	0.202
入場料	0.842	0.073	0.007	0.280	0.325	0.110	0.193	0.175

5．2　問いの再設定

以上，代理変数を用いて算出した各施設の特徴と外形的データである延べ床面積，飼育種類数，来場者数との相関を算出し，相関，あるいは逆相関の関係が強いものについて考察した（上記箇条書き➡）。これらの分析の結果（上記箇条書き●）はすべて仮説である。しかし，実務家がこれを用いる場合には，この結果を意思決定の際の根拠データとして用いることも可能である。

しかし，単語の意味を文脈なしに特定すること，つまり上記のような相関分析の結果だけでは十分に内容を把握することが難しいこともある。たとえば，「巨大」や「規模」，「大きい」と来場者数が相関関係にある一方で，延べ床面積と「最高」や「楽しむ」は逆相関の関係にあった。「not楽しむ」の施設の来場者数が多いとは思えないため，「巨大」や「規模」，「大きい」が延べ床面積（施設の広さ）を指しているとは考えられず，「巨大」，「規模」，「大きい」は水槽のサイズを指している可能性が高い（「水槽」と来場者数の相関も強いため）と推測される。このように，相関関係を算出しただけでは，その結果を十分に把握することは難しい。

このような把握の難しさを考慮し，次項ではさらなる分析を行う。特に前項最後の「来場者数は「夕方」と相関関係にある」を新しい問い，つまり

「なぜ「夕方」は来場者数と相関関係が強いのか」として再設定し，これにアプローチしていく。検証にあたっては，これまでの分散表現テキストマイニングに計量テキスト分析を併用する。その理由としては，分散表現テキストマイニングでは「文章を分散表現として要約してはいるが，分散表現は 100 次元（筆者注：ハイパーパラメーターの設定によって次元数は異なる）の座標で表現されており，人の目でこれら座標から中身を判断することは困難（竹岡，2018，p. 115)」であることがあげられる。

5．3　問いへのアプローチ

まず，全クチコミの中から「夕方」という単語が登場する文章（157 件）を抽出し，抽出された文章をK-means法で 2 つにクラスタリング（各クラスターに含まれるクチコミ件数は 78 件と 79 件）する。そのうえで各クラスター内での共起語を確認すると次のことがわかる。

第 1 クラスターに分類された文章の共起関係を確認すると，「夕方」と共起する単語として「水槽」が多く（20 件），さらに「夕方∩水槽」と共起する語としては「ジンベエザメ」が多い（10 件）ことが確認された。続いて，第 2 クラスターを確認すると，「夕方」と共起する単語として「チケット」が多く（21 件），「夕方∩チケット」と共起する語としては「行く」や「安い」が多い（各 12 件，10 件，重複あり）ことが確認された。以上から，仮説①夕方のジンベエザメの水槽が来場者数に影響している，仮説②夕方には割引チケットがあり，それが来場者数に影響している，の2仮説が立てられる。

これらの結果をもとに，仮説に対する解とその妥当性を検証するためには，アンケート調査等，さらなる調査が必要である。しかし，本稿の目的は水族館の来場者数に影響を与える要因の研究ではなく，分散表現を用いたテキストマイニングの方法の提案であるため，インターネット上の検索結果と既存のデータから，妥当性を確認することとする。まず，「沖縄美ら海水族館　夕方」をgoogleで検索すると確認できるwebページ[12]内には，16 時以降

12）https://churaumi.okinawa/topics/1500885803/（最終確認日：2018 年 10 月 18 日）

の入館には割引があること，17時にはジンベエザメの給餌があることの記述が確認される。また，「海遊館　夕方」でも，17時以降の「夜の海遊館」[13]に関する記述が確認され，またジンベエザメの写真も掲載されている。しかし，海遊館においては，割引チケットに関する記述は確認できない。名古屋港水族館に関しては期間限定の夕方割引チケットに関する記述が確認され，鴨川シーワールドと鳥羽水族館については夕方および，ジンベエザメ，割引チケットに関する記述は確認することができなかった。

さらに，来場者数と「夕方」，「ジンベエザメ」，「割引」の相関を確認すると，両概念ともに来場者数と強い相関関係[14]が確認される（表7）。また，各施設名と各概念の類似度を確認すると，「夕方」と「ジンベエザメ」については「沖縄美ら海水族館」と「海遊館」で高く，また割引についても「沖縄美ら海水族館」と高い類似度を確認することができる（表7斜字）。以上から仮説①「夕方のジンベエザメの水槽が来場者数に影響している」についてはその可能性が高いこと，②「夕方には割引チケットがあり，それが来場者数に影響している」についてもその傾向があることが確認される。

表7　検証に登場する概念と来場者数の相関

	相関	海遊館	沖縄美ら海水族館	鴨川シーワールド	鳥羽水族館	名古屋港水族館
来場者数（万人）		217	281	80	83	199
夕方	0.893	*0.286*	*0.307*	0.092	−0.035	0.136
ジンベイザメ	0.894	*0.448*	*0.584*	0.139	0.126	0.192
割引	0.811	0.156	*0.240*	0.146	0.137	0.160

6．分散表現テキストマイニングの内包する問題とそれに関する考察―再現性の観点から

以上2種類の分析を行ってきた。これらはともにfastTextによって算出された分散表現をもとにテキストマイニングを行ったものであるが，fastTextのようなニューラルネットワークを用いた手法を研究の方法として

13）https://www.kaiyukan.com/program/night/（最終確認日：2018年10月18日）
14）「割引」に関してはp値が0.09と大きく，相関関係があるとは言い切れない。

採用する際には，計量テキスト分析とは異なる問題点がある。それは再現性に関する問題である。再現性とは「誰もが，同じ手続きを用いて再び同じ測定ができ，最初の測定結果を追試できる可能性のこと（May，2001　邦訳書 p. 107）」で，ニューラルネットワークを用いる際には，これを担保することが難しい。

分散表現テキストマイニングを使用するにあたって，単語を分散表現化（ベクトルに変換）する際，fastTextでは重みづけに用いる行列を生成する部分でランダムに初期値を生成する処理が行われる（図 7 内 WinとWout）。そのため，全く同じテキストデータをfastTextで処理しても，毎回異なる結果が出力されることになる。ここでの結果とは，ひとつの単語に対して設定した次元でベクトル表現されたものを指す。たとえば，「テキスト」という単語を含むいくつかの文章をfastTextで学習させた場合，「テキスト」という単語に対して 1 回目は「(0. 52，0. 73，0. 05)」であったものが 2 回目には「(0. 35，0. 96，0. 76)」という結果が出力されるということである。このようにベクトル表現化した場合の数値が異なることで，これらを用いて計算する類似度も異なることになる。ある程度はepoch（学習回数）を多く設定することで解消可能な問題だが，結果を完全に一致させることは困難だと思

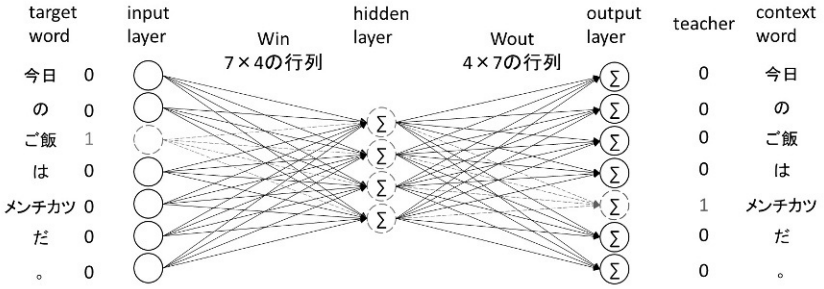

図 7　fastTextで用いるニューラルネットワーク図（簡略版）

われる。

同じデータを扱っても異なる結果が生じる，つまりその研究に再現性がないということになり，研究結果の信頼性を損なうことになりかねない。このような再現性の問題は，今後機械学習を含むニューラルネットワークを用いたAI技術が研究方法に取り込まれる中で，ひとつの論点になる可能性もある。以下では，この問題に関する考察を行うことで，今後の研究につなげていきたい。

研究の進め方は研究者や研究対象，そして選択する手法によって異なるが，テキストマイニングにおいては，対象の決定から，考察にかけて，およそ図8のようなプロセスで進むと思われる。この中で，分析対象の決定は再現性の対象にはならず，また，考察も再現性の対象にはならない。なぜなら，すでに決定されている分析対象について，その分析を再現するのが再現性であり，また，分析の結果あらわれた数値などをもとに何らかのインプリケーションを引き出すのが考察であるからだ。再現性の対象に含まれるのは，データの収集から分析までである。

図8　テキストマイニングのプロセス

以下が，各プロセスにおける再現のために必要な要素である。

1．データの収集
- 収集を行った日，場所，データの境界，方法など

2．クリーニング，整形，サンプリング
- 表記ゆれを修正する場合には「AをBに変換」などの情報
- 何（記号など）を，どの順番で削除したのか

3．分かち書き
- 通常何らかのソフトウェアを用いて行われるため，どのようなソフ

トウェアを使用したのかといった情報

4．集計と分散表現化

- 計量テキスト分析では単語ごとの集計が行われるだけなので特に問題はない
- 分散表現テキストマイニングでは単語のベクトル表現化が行われるが，ここで上述のランダムに生成される初期値を用いた計算がなされるため，再現性の問題が生じる

5．分析

- 再現に必要な情報として分析の詳細な手順など
- 初期値をランダムに生成するK-means法のような機械学習技術を用いる場合には再現性の問題が再び生じる

厳密な再現性があるとは，このすべてのプロセスを第三者が研究者と同一に行うと全く同じ結果が得られる状態を指す。分散表現テキストマイニングを行う際に再現性が問題となるのは，主に分散表現化のプロセスである。これは第三者の再現実験だけではなく，研究者自身が行っても，完全に一致する結果は得られない。

この問題に関する本稿の結論としては，その他のプロセスは通常のテキストマイニングと同一であり，問題となるのは分散表現化のプロセス，具体的にはこのプロセスで生成される学習済みモデルが試行のたびにごくわずかに変わることのみであること，しかし，分散表現化に使用した全単語およびその際のハイパーパラメーター（学習回数や学習率など，分析者が分散表現化に際して設定する値）は把握可能であり，それゆえに近似することは可能であること，つまりおおよその追試・検証は可能であり，また他者も同一の学習済みモデルを用いれば同じ分析の結果を得ることもできること，よってこの手法は，完全な厳密性を要求されるような分析でない場合には，十分に採用する価値のあるものであると考える。

7．おわりに

本稿では分散表現テキストマイニングを用いて仮説の検証を行った。この方法の利点は，web上のクチコミなど二次データを利用することができるのでコストを抑えることができ，また本研究のように（上位出現200語に絞ったものではあったが）総当たりで概念間の関係を調査することができるため，アンケートを用いた調査のようにあらかじめ項目を決める必要がなく，その恩恵として意外な結果が出ることも期待できる。また，今回の分析結果を見ると，かなりの程度うまく現実を写像しており，競合との相対的関係を可視化する道具としてはかなり有効といえる。とはいえ高木・竹岡（2018）が指摘するように，テキストマイニングは仮説検証型の研究よりも仮説構築型の研究に向いたものであり，前項で見たように，特に分散表現テキストマイニングは厳密な再現性を担保できないという点で，より仮説構築を志向する研究に適しているといえる。

しかし，本研究では仮説の構築とともに，仮説の検証も試みた。その結果の妥当性の検証は必要だが，控えめにも，分析者が本格的な分析を行う前のパイロットスタディとしては有用であることは確かである。

分散表現テキストマイニングは誕生したばかりの技術であり，その使用法についても確立されてはいない。今後もこの方法の応用について検討していきたい。

謝辞

本研究はJSPS科研費，若手研究(B)JP17K13787「イノベーションの普及過程で選好される意味属性のテキストマイニングによる可視化」の助成のもとなされたものです。

参考文献

鈴木潤，海野裕也，坪井祐太（2017）「言語処理における深層学習の基礎」坪井祐太，海野裕也，鈴木潤『深層学習による自然言語処理』講談社，pp. 43−90.

高木修一，竹岡志朗（2018）「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―」『富大経済論集』第 64 巻 2 号，印刷中.

竹岡志朗（2018）「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・高木修一（2018）「wwwにおけるクチコミ情報収集の方法に関する考察―人の情報探索行動の観点から―」『経営研究』第 69 巻 1 号，pp. 91−107.

樋口耕一（2014）『社会調査のための計量テキスト分析』ナカニシヤ出版.

Bengio,Y., Ducharme, R., Vincent, P., Jauvin, C.（2003）“A neural probabilistic language model,” Journal of machine learning research, 3, pp.1137−1155.

Harris Z.（1954）“Distributional structure,” *Word* Vol. 10 , No. 23, pp. 146−162.

Joulin, A. Grave, E., Bojanowski, P., Mikolov, T.（2016）. ”Bag of tricks for efficient text classification,” arXiv preprint, arXiv：1607.01759.

May, T.(2001) *Social Reseach 3 rd edition,* Open Univesity Press（中野正大監訳（2005）『社会調査の考え方　論点と方法』世界思想社）.

Mikolov, T., Yih, W. T., Zweig, G.（2013）“Linguistic regularities in continuous space word representations,” *Proceedings of Naacl-HLT* 2013, pp. 746−751.

Rubenstein, H., & Goodenough, J. B.（1965） “Contextual correlates of synonymy,” *Communications of the ACM,* 8(10), pp.627−633.

（たけおか・しろう／本学兼任講師／2018 年 11 月 7 日受理）

Text Mining Using Machine Learning（2）
―― Discovery and Verification of Hypotheses ――

TAKEOKA Shiro

Abstract

In recent years, AI has gained social attention as there are reports of leap in image recognition contest ILSVRC, the appearance of AlphaGo, and applied cases in various industrial fields. Words related to AI such as machine learning and deep learning has become widely recognized. As an application possibility to management research and management practice, it can be said that it is becoming a sufficiently usable technology due to improvement of natural language processing technology in recent years, especially improvement of distributed representation technology.

In this research, we propose a method to analyze and visualize the characteristics of goods and services with text mining which using machine learning technology included in AI technologies, especially the method on discovery and verification of hypotheses. The text mining method proposed in this research is not based on current weighing text analysis which aggregate values of words, but based on distributed representation of words calculated by machine learning. By using this method, it is possible to analyze based on the consumer's experience and meaning made up of the consuming process. The method proposed by this paper is also useful for business researchers, but even for practitioners, when planning and model change of new products or services, comparison with other company's product services becomes easier than ever, more detailed analysis It is thought that it is beneficial because it can be considered to be able to proceed with practical work based on it. Text mining based on distributed representation is still an incomplete technology, not a fixed method, but it is an area where further development is expected in the future.