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 ――

竹　岡　志　朗

1　はじめに

近年，google社の開発したAlphaGoの性能が囲碁の世界だけではなく社会的にも話題となり，機械学習やディープラーニングといったAI（artificial intelligence：人工知能）に関する言葉が広く認知されるようになってきた。これらの技術の進歩は早く，現在でも画像認識技術を活用した医療補助AIについての議論が活発になるなど，多方面での応用が期待されている。経営学研究および経営実践への応用可能性としては，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。

また，ポーターの5force分析における代替品の脅威にあたる，スマートフォンとデジタルカメラやwww（world wide web）と書籍のように直接的な競合関係，つまり同一の商品・サービスカテゴリー内での競合関係ではなく，商品カテゴリーを超えた競合関係が話題にもなっている。

本稿では，以上の近年の社会的関心のもと，機械学習技術を用いて商品・サービスのカテゴリーを超えた競合関係をテキストマイニングによって分析・可視化する手法を提案する。本稿が採用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といった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いたものである。そのため，分析の基盤はテキスト内の単語の使用頻度ではなく，単語の分散表現を用いた類似度となっている。分散表現に基づいたテキストマイニングは，ま

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

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。本稿では，この分散表現を用いた 2 つの分析手法を提案する。

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，有益だと考える。

2　計量テキスト分析

現在のテキストマイニングの主流は計量テキスト分析と呼ばれるものである。計量テキスト分析とは樋口（2014）によれば「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（content　analysis）を行う方法（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を集計し，その集計値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いたりすることである。これらの手法は経営学だけではなく，看護学や教育学などでも広く用いられており，様々な質的データの分析を定量的に行うことを可能にしている。また樋口（2013）のように，アンケートで収集された属性データと自由記述データを組み合わせた統計的分析などもある。

しかし，齋藤（2011）は，現在のテキストマイニングは応用事例が多いものの分析という観点から見ると手法が単純集計，共起率の分析，テキストの持つ属性の出現単語からの分析といった記述的なものが大半を占めており，推測的な手法を用いたものが少ない，と指摘している。その原因としては，テキストマイニングという手法が応用研究を行うものにとっては難易度が高く，その分析をソフトウェアの手順に従ったものに留めている点を挙げている。

また，基本的な計量テキスト分析は単語や文章の持つ意味への接近に弱み

がある。この問題への対策として次項の手法がとられている。

3　意味へのアプローチ

計量テキスト分析では，同じ単語であってもその意味は文脈に依存して決まるため，意味に接近することが困難である。そこで意味への接近手法として共起関係の分析が行われる。「共起する概念の数が多いほど解釈の余地が狭まり，概念に充当されている意味の共通性が近似する可能性が高くなる（竹岡，2016a，p. 105）」のである。樋口（2011）や竹岡（2016b）では，共起関係に関する分析を発展させ，複数語による組み合わせでコーディングルールを作成し，分析対象となる文章の意味を分析している。

たとえばコーディングルールとしては

「メモリカード」 ∩ {「ファイル」 ∪ 「写真・動画」} ∩ 保存

のように出現単語の組み合わせを設定し，これに対して「メモリカード保存」というラベルを張る。つまり，メモリカードとファイルあるいは写真・動画，そして保存という概念が文章中に入っていれば，「メモリカード保存」についての話題に関する文章であると推定する。ここで単語ではなく概念としたのは，「メモリカード」にはメモリカード，SD，SDカード，ミニSD，マイクロSD，メモリスティックという単語が含まれるためである。このように頻出する類語に関しては概念としてまとめることで，より詳細な意味の抽出が可能になる。

しかし，このようなアプローチによる文章の持つ意味への接近には限界もある。そもそも計量テキスト分析は後述のone-hotベクトル表現のようなもので，すべての文意を抽出するためには巨大かつ疎なベクトルデータを用いる必要があり，計算数が出現単語数に合わせて増加するという問題がある[1)]。そこで，樋口（2011）や竹岡（2016b）では，コーディングルールに使用する単語を頻出語などに限定したうえで，頻出共起関係を用いてコーディング

1）たとえば，1万語の異なり語が登場するクチコミを分析する場合には，1万次元のベクトルを用いる必要がある。

ルールを作成している。

しかし，このような手法では，コーディングルールとして作成されたラベルの数が少なければ頻出共起関係だけを分析の対象とすることになり，多くの文章が分析の対象外になりかねない。これに対応するためにはコーディングルールを相当数作成する必要があるが，これによって計算数の増大を招くことになる。このような問題に対するひとつの解決法がベクトルの次元を数百程度に圧縮し分析することを可能にする分散表現を用いたテキストマイニングである。

4　fasttext[2] を用いた分散表現（distributed representation）分析

4．1　分散表現

本稿が提案する機械学習を活用したテキストマイニングでは，これまでの計量テキスト分析では出現回数や共起関係によってあらわしていた単語や文章の特徴を，分散表現を用いてあらわすことになる。

分散表現とは「任意の離散オブジェクトの集合Vに対して，各離散オブジェクトv∈Vにそれぞれd次元のベクトルを割り当て，離散オブジェクトをD次元ベクトルで表現したもの（鈴木他，2017，p. 58）」とされ，また離散オブジェクトとは「人や物の名前，概念のように，物理的に計測できる量を伴わず，通常，記号を用いて離散的に表現するもの（同上）」である。テキストマイニングの文脈で分散表現を平易に表現すれば，分析対象としてある各文章や各単語を決められた次元のベクトルで表現したものといえる。

文章や単語をベクトル化する方法には，局所表現と呼ばれるベクトルのすべての要素の中である要素のみが1，その他が0で表現されたone-hotベクト

2）本稿では機械学習を行うアプリケーションソフトウェアとしてFacebook社が開発したオープンソフトウェアfasttextを使用する。fasttextと同様に分散表現を用いた機械学習ソフトウェアとしてはWord2vec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。プログラミング言語pythonなどを使用すれば，複数のモジュールがすでに用意されており，容易に使用することができる。また，ターミナルを使用すれば直接fasttextで分析することも可能である。

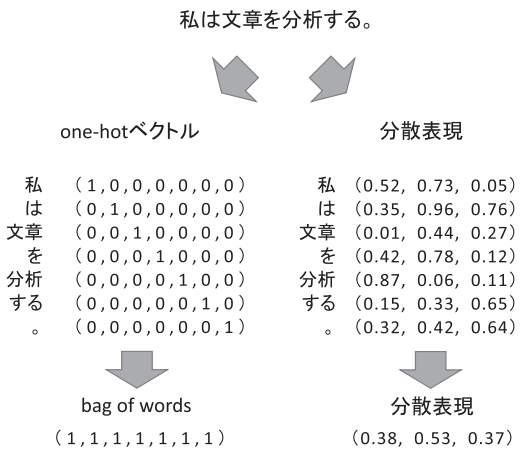

図1　文章のベクトル化― one-hotベクトルと分散表現

ル表現もあるが，これは分析対象となるすべての文章に登場する異なり語数がベクトルの次元数となるため，巨大かつ疎な行列を計算することになる。この場合，文章はbag of wordsの形式でベクトル表現される。

他方，分散表現では，単語のベクトル化にはSkip-gram[3]やCBOW（continuous bag of words）といった手法がとられ，これらによって各単語のベクトルが算出される。これらの手法を用いることによって各単語のベクトルの次元数を通常で100次元，wikipediaの全データを用いた場合でも各単語を300次元で表現することが可能となり，局所表現と比べて大幅に次元数を圧縮することができる[4]。

3）本稿では，文章分類を行うことから，分散表現の算出に教師あり（supervised）を使用している。単語の分散表現を得るだけであれば，チュートリアルではSkip-gramが推奨されている。

4）fasttextでは，分析者がベクトルの次元数を任意に設定することができる（1から1000まで確認）。日本語版チュートリアルではベクトルの次元数設定について，50M tokens：100次元，200M tokens：200次元，最大：300次元とされている。（https：//github.com/icoxfog417/fastTextJapaneseTutorial：2017年11月28日確認）また，同チュートリアルではtokensについて単語数ではないかとされているが，Mikolov & Sutskever et al.（2013）の中での表現では句，つまり“Boston Globe”という新聞名のように2語以上の語を合わせることでひとつの意味を成

4．2　分散表現における意味へのアプローチ

計量テキスト分析においては，単語や文章の意味を特定することが難しかった。分散表現を用いた分析においては単語をベクトル表現に変換する過程で単語に文脈情報を持たすことができる。たとえば，図 2 左のような離散表現の場合，出現回数を集計しただけでは単語間の類似度を測ることができない。しかし，同図右のようにSkip-gramやCBOWを用いてベクトル表現に変換すれば単語間の類似度をコサイン類似度[5]で測定することができる。

このようなことを可能にしているのがSkip-gram[6]やCBOWの背景にある

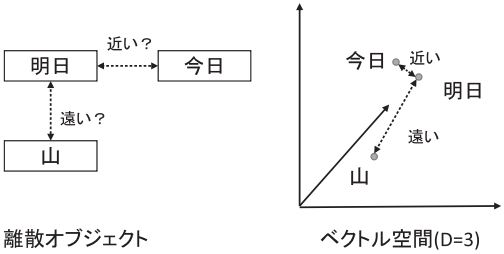

図 2　離散オブジェクトとベクトル空間
（鈴木他, 2017, p. 59　図 3. 2 を一部改変）

すものをtokenとして扱っており，このことから文章表現の中で意味を成す最小構成要素をtokenとしていると考えられる。日本語を分析する場合にはtokenという考え方は必要ないだろう。なぜなら，fasttextのような分散表現分析だけではなく，これまでの計量テキスト分析においても，テキストを分析する際には文章の単語間を半角スペースで区切った分かち書き済みのテキストファイルを投入するが，日本語では通常の文章の場合，英語などのように単語間にスペースは存在せず，形態素解析ソフトウェアを用いることで初めて分析可能なテキストファイルにすることができる。ゆえにtokenは，日本語の場合，形態素解析を行う際に「携帯」と「電話」ではなく「携帯電話」というように意味を構成する最小構成要素になるように分かち書きしておけば考慮する必要がないためwords＝tokensと考えればよい。このtokenについては英語などを分析する際には考慮する必要のあるものといえる。

5）コサイン類似度は次の式によって計算することができる。$\cos(\mathrm{A},\mathrm{B}) = \dfrac{\mathrm{A} \times \mathrm{B}}{|A||B|}$

6）本稿が使用するソフトウェアfasttextにおいてSkip-gramを用いて分散表現に変換する際には，subword information（部分語情報）も活用されている。Bojanowski et al.（2017）は，subword informationを用いることで，まれにしか出現することのないような単語でも信頼できる分散表現を得ることができるとしている。

分布仮説（Harris，1954，Rubenstein & Goodenough，1965）である。分布仮説とは類似する文脈で出現する単語は，意味的にも類似しているという考えであり，Skip-gramやCBOW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。ゆえに図2右のように「今日」と「明日」という類似する文脈で出現する可能性の高い単語はベクトル空間上で近く，「明日」と「山」のように同一文章で登場する可能性は高いが，類似する文脈で登場する可能性の低い単語はベクトル空間上では遠くなる。

先述の通り，分散表現では各単語をSkip-gramなどの手法でベクトル表現に変換することで，単語間の類似度をコサイン類似度として計算することができ，また，単語間の関係をベクトルのオフセットで推定することも可能となる（図3は二つの単語の単数形と複数形の関係を図示したもの）。

このような分散表現の最大の特徴は，加法構成性を持つベクトルの加減算による類推にも応用することができる点にある。たとえば図4のように「King」−「Man」＋「Woman」は「Queen」に近似する（Mikolov & yin et al.，2013）といった類推が可能となる。このような計算が可能ということは，この手法がかなりの程度正確に単語の意味をベクトル表現として把握していることを意味している。

以上のように，単語の分散表現では単語間の関係をコサイン類似度で算出

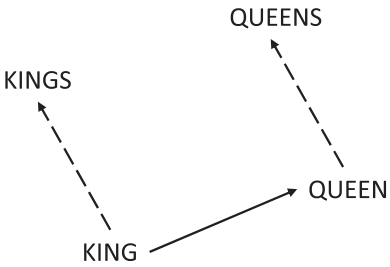

図3　単語のオフセット
（Mikolov et al., 2013, p. 749 figure2を一部改変）

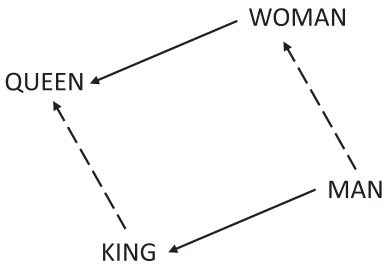

図4　分散表現の加法構成性

したり，ベクトルの加法構成性を利用して意味の類推を行うことができる。なお，文章のベクトル化[7]には，本稿で使用するfacebook社が開発した自然言語処理に関する機械学習ソフトウェアfasttextの場合には，文章を構成する単語のベクトルとEOS（end of sentence）「＜/s＞」のベクトルを合計した後に文章を構成する単語数＋1（この＋1 はEOS）で除したものを用いている[8]。このように計算することですべての文章がベクトル表現化され，計量テキスト分析では困難だったすべての文章の持つ意味を分析の対象にすることができる。

5　機械学習を用いたテキストマイニング

5．1　分析の対象と環境

以下では，www上のクチコミデータを用いて機械学習を活用したテキストマイニングを行っていく。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www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www上

7）文章を分散表現に変換するソフトウェアとしては，fasttext以外にもDoc2Vecなどがある。

8）文章のベクトル計算については，単語のベクトルを正規化した後に合計し，それを単語数で除するという記述がwww上では散見されるが，筆者の環境下ではベクトルを正規化しない上記式によって算出することができた。

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（2017）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。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り，このような不均衡データを用いたクラス分類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1000件をサンプリングし，計20000件である。

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。

fasttextはデフォルト設定で使用した。

分析に使用した環境は下記のとおりである。

OS	ubuntu 16.04.3 LTS
形態素解析	mecab 0.996
辞書	mecab-ipadic-neologd　2.7.0
機械学習	fasttext“431c9e2a9b5149369cc60fb9f5beba58dcf8ca17”[9]
プログラミング言語	python 3.5.2
使用モジュール	pyfasttext 0.4.4[10]（fasttextのモデルに接続） scikit-learn 0.19.1（k-meansによるクラスタリングに使用）

5．2　分散表現（distributed representation）の分散（variance）に基づく知覚マップの作製

分散表現を用いれば，図2右にあるように，意味の類似する単語をコサイ

9）fasttextはバージョン番号が公表されておらず，ソフトウェアのhash値からバージョンの違いを確認するとされている。（https：//github.com/vrasneur/pyfasttext：2017年11月28日確認）

10）www上ではgensimを用いてfasttextのモデルに接続する例がよく見られるが，筆者の環境下ではgensimによる算出結果とfasttextで直接行った計算の結果が一致しなかった。そのため本稿ではpyfasttextをfasttextに接続するモジュールとして使用している。

ン類似度を算出することで抽出することができる。コサイン類似度は 1 から −1 の間の値を取り，1 に近いほど類似度が高いことを意味している。たとえば今回使用しているデータのなかで東京ディズニーランドといくつかの施設のコサイン類似度を算出すると，東京ディズニーシーが 0.73436，ユニバーサル・スタジオ・ジャパンが 0.41962，三井アウトレットパーク木更津が −0.58749 である。

これを応用すれば各施設を構成する要素とその構成の程度の近似値を算出することができる。たとえば，「東京ディズニーランド」と「パレード」の類似度は 0.994496，「アトラクション」の類似度は 0.923629 であり，「ショッピング」の類似度は −0.58031 である。この結果から東京ディズニーランドはパレードやアトラクションがその特徴として消費者に認識されてお

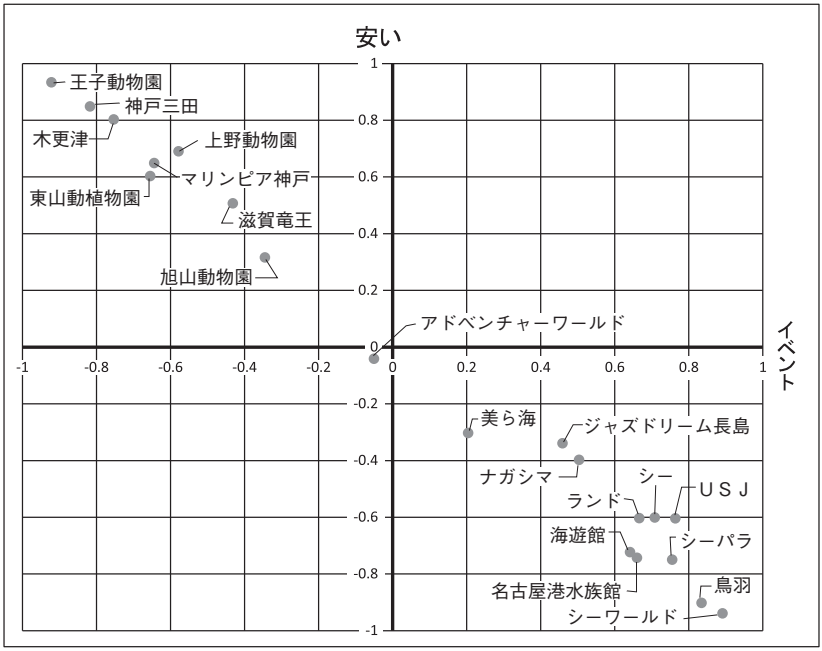

図 5　「イベント」と「安い」を 2 軸にとった知覚マップ[11)]

11）グラフの可読性を考慮しそれぞれの施設名は略語を使用している。

り，ショッピングを行うところとは認識されていないことが推測される。このコサイン類似度を基に作成した知覚マップが図5である。

この知覚マップの作成に当たっては，クチコミ中の名詞，動詞，形容詞を合わせた上位頻出200語を対象に，それらの語と各施設名のコサイン類似度を算出，それらの分散を計算し，分散の大きなもの（「安い」の分散は0.425845421，「イベント」の分散は0.416727697）を二つの軸にとって各施設をプロットしている。たとえば，グラフ中左上にある王子動物園は「神戸市立王子動物園」と「安い」のコサイン類似度が0.933182063と高く，「イベント」とのコサイン類似度は−0.92163741と低いことをあらわしている。

図6は図5と同様に分散の大きい「お土産（分散：0.377282719）」と「楽しかっ[12]（分散：0.417025259）」を2軸にとり，それら2単語と各施設名の

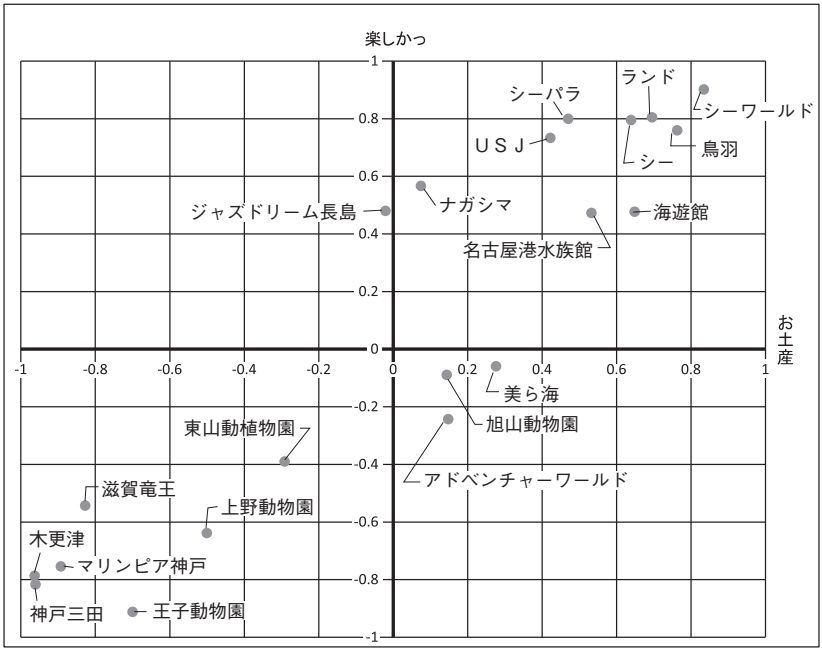

図6　「お土産」と「楽しかっ」を2軸にとった知覚マップ

12）脚注15を参照。

類似度をもとに，各施設名をプロットしたものである。このグラフからショッピングモールの多くはお土産との類似度が低いことや，「お土産」との類似度が高い施設ほど「楽しかっ」との類似度が高いことがわかる。

このように，分散表現とそれに基づく類似度を用いれば知覚マップを容易に作成することが可能である。また，アンケートによるデータ収集のように最初に決められた概念に関する集計結果を用いて作成するのとは異なり，何らかの形で単語がクチコミなどのデータ内に出現していれば，このような形で知覚マップを作成することが可能である。

また，今回は知覚マップ中に広く分布するように分散の大きな単語を用いたが，極論すれば，データ中に登場するすべての単語をこのような形で分析することも可能であり，これまでと比較して，容易に素早く消費者の観点から見た分析対象の市場におけるポジショニングを確認することができる。

5．3　文章分類に基づくカテゴリー横断分析

次に，自然言語を対象とした機械学習の中でも広く利用されているテキスト分類の機能と，これまでの計量テキスト分析の手法を用いて各施設に共通して言及されている特徴を可視化する。

機械学習におけるテキストの分類は教師あり学習で行われることが多く，各テキストにラベルを張り，それをもとにテキストの特徴を抽出，学習するというプロセスがとられる。そして，このプロセスで学習された結果をもとにして，新たにテキストが投入された場合には，そのテキストがどのラベルかを判定し，分類することが可能になる。通常はこのように用いられるのだが，本稿では学習に使用したデータを再度fasttextに投入し，分類器によって判定，その中で分類器が誤判定したラベルから，各施設に共通している特徴について分析を行う。

表1では，分類器の判定結果が元のラベルと一致する場合には一致，ラベルは一致しなかったが，同一カテゴリーの施設のラベルが張られた場合には同一カテゴリー，カテゴリーも一致しなかった場合には不一致として，それ

表 1　分類器による再ラベリングの結果

	一致	同一カテゴリー	不一致
ディズニーランド	7	894	99
ディズニーシー		930	70
ユニバーサル・スタジオ・ジャパン	837	99	64
横浜八景島シーパラダイス	631	262	107
ナガシマスパーランド	401	505	94
沖縄美ら海水族館		52	948
鳥羽水族館	7	14	979
鴨川シーワールド	44	10	946
海遊館		19	981
名古屋港水族館	8	12	980
アドベンチャーワールド	115	660	225
旭川市旭山動物園	727	170	103
神戸市立王子動物園	174	718	108
東山動植物園	539	296	165
上野動物園		882	118
神戸三田プレミアムアウトレット		976	24
三井アウトレットパークジャズドリーム長島		927	73
三井アウトレットパーク木更津		965	35
三井アウトレットパーク滋賀竜王	966		34
三井アウトレットパークマリンピア神戸		955	45

らの結果を集計している。

表 1 の分類器による再ラベリングの結果から，「ユニバーサル・スタジオ・ジャパン」や「三井アウトレットパーク滋賀竜王」のように，分類器がかなりの程度判定に成功しているクチコミもあれば，「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」の各水族館のように，他のカテゴリーの施設と判定されているものもある（脚注 14 も参照）。このような結果からは，「ユニバーサル・スタジオ・ジャパン」や「三井アウトレットパーク滋賀竜王」の場合には，他の施設と比較して何らかの特徴的なものが存在し，それが分類器の判定を正確なものとしていることが推測される。

以下では不一致カテゴリーに注目する。このカテゴリーに含まれているクチコミは先述のとおり分類器によってラベリングした際に，異なるカテゴリーのサービス施設名がラベリングされたクチコミである。このようなクチ

コミに注目する理由としては，このようなクチコミにはサービスの業態を超えた共通点があると考えられるからである。つまり，ラベルが一致したクチコミを分析すればその施設の特徴が，同一カテゴリーに含まれたクチコミを分析すればそのサービスカテゴリーの特徴が，そして不一致カテゴリーを分析すれば今回の分析対象となった 4 つのサービスカテゴリーに共通する特徴が抽出できると考えられる。

しかし，このラベルが不一致のクチコミは全体で 6198 件存在する。これらを直接分析することは困難である。そこで，これらのクチコミを共通の特徴に沿って複数のクラスターに分割し，それを分析することで，より消費者の意見を可視化することが容易になる。先ほどのfasttextを用いたクラスタリングでは教師データを用いてクラスタリングを行ったが，今回はこれらのラベル（施設名）を用いても意味はない。なぜなら，クチコミを施設ごとに分類するのが目的ではなく，クチコミ内に登場する何らかの特徴に基づいて分類することが目的となるからである。そのため，教師データのない状態でクラスタリングを行う必要がある。

本稿では教師データなしでのクラスタリングにk-means法を用いる。k-means法ではfasttextが算出した文章の分散表現を用いてクラスタリングを行った。図 7 のエルボー図より，クラスター数は 8 から 15 が適正と推測し，

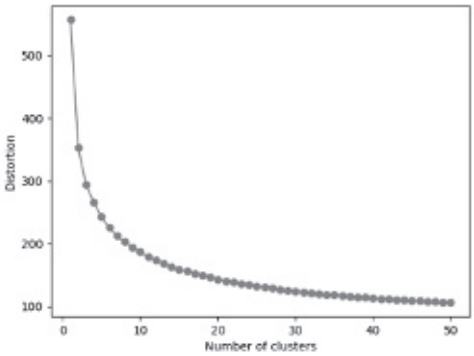

図 7　エルボー図

今回は全クラスター内のクチコミ数が1000以下となる分岐点，11クラスターに分類することとした。

ここまでで共通話題のクラスタリングはできたが，この各クラスターの中でどのような内容が書き込まれているのかを分析する必要がある。しかし，fasttext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である。そこで，以下では計量テキスト分析の手法を用いて，各クラスターの中でどのような話題が投稿されているのかを見ていくことにする。

図8は各クラスターの頻出語（名詞，動詞[13]，形容詞）と出現回数を示している。図中の各表の左上にはクラスター名を，右上には各クラスターに分類されたクチコミ件数を，第1列は頻出語上位3語を，第2列にはその出現回数を，第3列には各語との共起頻度の高い上位3語を，そして第4列には共起語の共起数を記している。たとえばクラスター0の出現回数1位は「ショー」で全655件のクチコミ中420件で登場しており，この「ショー」と共起する形で「見」が153件登場している。

この結果から，単語の出現数には偏りがあること，つまり各クラスターの特徴は抽出されていることがわかる。たとえば，クラスター1は大きな水槽について，クラスター2はパンダについて，クラスター0や5，9は水族館でのショーについて，クラスター8は駐車場について，クラスター10は子供を連れての行き先として言及されていることが推測される[14]。

13）「ある」，「いる」等の意味抽出の困難なものは除いている。

14）図8を見るとクラスター0，1，3，4，5，9は人であれば水族館に分類するような単語の出現の傾向であり，11クラスター中の半数の6クラスターがこれにあたる。表1では水族館の各施設が不一致カテゴリーに分類される件数が多かったが，その結果の偏りがこのようなクラスターの偏りを生んだと考えられる。このようにfasttextが適切に文章を分類できなかった理由としては，ラベル数が20と多いこと，またかなり類似する施設を分類していることなどがあげられる。たとえば，「良い」と「悪い」という評価を学習させて，それをもとに分類させれば，そこで出現する単語の分布の違いは明確であり，今回とは異なる結果になると考えられる。また，水族館という施設が他のサービス施設と類似するサービスを提供しており，それへの言及への多さからこのような結果になったとも考えられる。

クラスター0　655件

ショー	420	見	153
		イルカ	142
		シャチ	138
水族館	235	ショー	112
		見	75
		行っ	60
イルカ	199	ショー	142
		シャチ	88
		見	77

クラスター1　594件

水槽	446	ジンベイザメ	182
		大きな	140
		見	125
ジンベイザメ	238	水槽	182
		大きな	67
		見	62
水族館	164	水槽	112
		見	55
		ジンベイザメ	49

クラスター2　115件

パンダ	84	行っ	16
		弁当	14
		平日	14
動物	26	見	6
		パンダ	5
		思い	5
		行っ	5
弁当	18	パンダ	5
		食べる	14
		持っ	98

クラスター3　267件

ショー	246	シャチ	164
		イルカ	129
		水族館	108
シャチ	178	ショー	164
		イルカ	101
		水族館	79
イルカ	137	ショー	129
		シャチ	101
		水族館	59

クラスター4　361件

水族館	249	水槽	141
		ジンベイザメ	79
		魚	58
水槽	240	水族館	141
		ジンベイザメ	87
		魚	76
ジンベイザメ	118	水槽	87
		水族館	79
		迫力	39

クラスター5　658件

水族館	346	思い	76
		行っ	69
		行き	68
ショー	136	水族館	34
		思い	31
		イルカ	31
楽しめ	126	水族館	59
		イルカショー	28
		ショー	27

クラスター6　657件

思い	141	行き	33
		行っ	33
		見	27
見	131	時間	29
		思い	27
		行っ	27
行っ	124	思い	33
		見	27
		行き	24

クラスター7　895件

見	280	ショー	100
		行っ	87
		行き	80
行っ	253	ショー	78
		行き	70
		思い	68
ショー	246	見	100
		思い	81
		行っ	78

クラスター8　478件

駐車場	91	行く	17
		車	16
		無料	16
思い	75	楽しめる	15
		駐車場	14
		良い	14
行き	69	思い	12
		駐車場	11
		アウトレット	11

クラスター9　587件

ショー	482	シャチ	219
		イルカ	168
		水族館	141
シャチ	270	ショー	219
		イルカ	88
		見	85
水族館	207	ショー	141
		シャチ	56
		行っ	54

クラスター10　931件

行き	278	行っ	62
		思い	61
		子供	60
行っ	232	行き	62
		子供	40
		思い	40
子供	161	行き	60
		行っ	40
		楽しめる	40

図8
各クラスターの
頻出語と共起語[15)]

15）表中では「見」や「行っ」のように活用語尾の無い動詞が登場している。例えば計量テキスト分析用ソフトウェアとして多方面で活用されているKH Coderの場合にはこのような語は基本形に直したうえで分析に使用しているので，これらの語は「見る」と「行く」として集計されていると考えられるが，本稿ではその後の分析に使用することなども考慮し，基本形には変換せず使用している。そのため，同じ「見る」という単語であっても本稿では「見」，「見る」，「見れ」といった形で登場している。なお，形態素解析に使用したmecabの辞書には基本形が登録されており，形態素解析を行う際に簡単に変換することもできる。

しかし，これらの特徴はそのクラスターの最大公約数的な特徴であり，細かな特徴をつかみきることはできていない。たとえば，クラスター 7 では，車いすに関する言及が 5 回，ベビーカーに関する言及が 30 回，アレルギーに関する言及が 2 回あり，またクラスター 10 では観覧車やサンタマリア，明石海峡，繁華街など近隣施設への言及と思われる単語が複数登場している。これらはそれぞれのクチコミを人の目で確認することで抽出可能な特徴といえる。

以上，2 種類の機械学習を用いた分析手法を提案した。これらの方法を用いれば，商品カテゴリー・サービスを超えた分析が可能となり，また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いため，効率の良い有用な手段だといえる。しかし，クチコミというデータには，それが消費者の意見のごく一部でしかないという問題点も存在する。つまり，クチコミを投稿するのはその商品やサービスの利用者のごく一部であり，またクチコミを投稿する際には，その商品やサービスにおける経験を振り返り，そのすべてを書き込むのではなく，何らかの基準で選択した結果を書き込むことになる。ゆえに，クチコミを分析した結果だけを結論を導き出す道具とすることは控えるべきだろう。しかし，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大変有用である。これらの手法を用いて可視化された特徴をもとにアンケートを作成し，それを分析すれば，これまで以上に効率よく，様々なことが分析可能と思われる。

6　計量テキスト分析と分散表現を用いた分析に関する考察

今回は機械学習という最新の技術を用いた分析手法の提案と，その応用および適用可能性の確認が目的であり，分析対象として「テーマパーク・レジャーランド」，「水族館」，「動物園・植物園」「アウトレットモール」という「休日のちょっとしたお出かけ先」の 4 つのカテゴリーを選んだが，この選択に特に意味はない。ある程度異なるサービスを提供しているものを選ぶことで可能性を検討したという以上のものはない。ゆえに，以下では分析結

果に関してはインプリケーションを考察せず，テキストマイニングという手法についての考察にとどめたいと思う。

分散表現を用いたテキストマイニングと計量テキスト分析を用いたテキストマイニングの相違点としては，計量テキスト分析が単語の出現回数に基づいて共起分析や階層的クラスター分析を行うのに対して，分散表現を用いたテキスト分析では単語や文章のベクトルを加法構成性に基づく加減算やコサイン類似度によって分析を行う点にある。

単語や文章を分散表現にすることで，計量テキスト分析では困難であった単語や文章の持つ意味への接近や，類推，分類が可能になる一方で，人の認識可能な特徴のかなりの部分が失われることにもなる。ゆえに，文章分類に基づくカテゴリー横断分析においては，分類器によって分類された何らかの特徴があると考えられる文章クラスターに対して，従来の計量テキスト分析の中でも最もシンプルな頻度分析を行うことによって，分類器の結果だけではわからない文章クラスターの特徴を抽出した。しかし，クラスターの持つ特徴を計量テキスト分析するだけでは見落としてしまうものも存在する。なぜなら計量テキスト分析，特に出現頻度に基づく分析は大きな特徴の抽出には向いているが，ごく少数しか現れない特徴の抽出は苦手としているからである。このような特徴は，共起ネットワークや階層的クラスター分析では足切りされ，比較的多数の単語を分析の対象として使用することのできる自己組織化マップ（樋口，2014）においても，分析対象になりうるかどうかはコンピューターの性能に依存するところがある。

他方で，このような文章クラスターの特徴の抽出に関しては，これまでの計量テキスト分析ではできなかった分類を，文章の分散表現への変換とそのクラスター化によって可能となっている。

今後ますます発展すると思われる機械学習の技術は，これまでには想像のできなかった分析手法を我々に提供し，これまでとは異なる分析，あるいはこれまでと同じ分析であったとしても異なるルートから結果に到達することを可能にすると思われる。しかし，機械学習を用いたテキストマイニングも

本稿で見てきたように万能ではない。ゆえに，今後は機械学習によるテキストマイニングについての応用を考えるとともに，これまでの計量テキスト分析との相互補完的な併用についても検討する必要があると考える。

最後に，機械学習を用いたテキストマイニングという手法だけに限定しても加法構成性や，類似度の分散に基づく分析は多様な応用が考えられる。たとえば「海遊館が東京ディズニーランドになるために必要な要素は何か？」という一見馬鹿げた考察も「東京ディズニーランド」のベクトルと，「海遊館」と単語Aの合成ベクトルとの類似度を計算することで明らかにすることが可能だと思われる。また，施設名と単語の類似度を計算することができることから，施設名を単語に代表させ，単語との類似度の分散を算出し，共通する特徴（誘因や課題），あるいは突出した特徴を可視化することも可能だと思われる。今後は以上の分析手法の構築を当面の課題としたい。

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

齋藤朗宏（2011）「日本におけるテキストマイニングの応用」『北九州大学ワーキングペーパーシリーズ』.

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』pp. 43–90　講談社.

竹岡志朗（2016a）「普及過程で連続的に変化する意味―テキストマイニングにおける三角測量分析―」竹岡志朗，井上祐輔，高木修一，高柳直弥『イノベーションの普及過程の可視化　テキストマイニングを用いたクチコミ分析』pp. 104–125　日科技連出版社.

――――（2016b）「イノベーションの普及過程における非連続性と連続性―テキスト

マイニングにおける話題分析―」竹岡志朗，井上祐輔，高木修一，高柳直弥『イノベーションの普及過程の可視化　テキストマイニングを用いたクチコミ分析』pp. 126−142　日科技連出版社.

竹岡志朗・高木修一（2017）「インターネットを用いた情報探索に関する検索エンジンとwebリンクの観点からの考察」日本情報経営学会第 75 回大会発表資料.

樋口耕一（2011）「現代における全国紙の内容分析の有効性―社会意識の探索はどこまで可能か―」『行動計量学』Vol. 38−1　pp. 1−12.

――――（2013）「情報化イノベーションの採用と富の有無：ウェブの普及過程における規定構造の変化から」『ソシオロジ』Vol. 57−3　pp. 39−55.

――――（2014）『社会調査のための計量テキスト分析』ナカニシヤ出版.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T.（2017）“Enriching Word Vectors with Subword Information,” in *Transactions of the Association for Computational Linguistics,* Vol. 5, pp. 135−146.

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

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J.（2013）“Distributed representations of words and phrases and their compositionality,” in *Advances in neural information processing systems,* pp. 3111−3119.

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

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

（たけおか・しろう／本学兼任講師／2017 年 12 月 5 日受理）

Text Mining using Machine Learning
—— Cross-sectional Analysis of Products／Service Categories using Reviews ——

TAKEOKA Shiro

Abstract

In recent years, AlphaGo developed by google has become a topic not only in the world of Go but also in society, words related to AI such as machine learning and deep learning are becoming widely recognized. Advances in these technologies are quick, and discussion about medical assistance AI which applies image recognition technology is active, and it is expected to be applied in various fields even now. As an application possibility to management research and management practice, improvement of natural language processing technology in recent years, especially progressive technology concerning distributed representation, can be said to be a fully usable technology.

And, it is not a direct competition relationship like a smartphone and a digital camera, www (world wide web) and a book, which is a threat of substitute in Porter's 5force analysis, that is, competitive relations beyond product categories are topics.

In this paper, we consider a recent social change and propose a method to analyze and visualize competing relationships beyond the categories of goods and services by machine mining using machine learning (AI) technology. The method of text mining adopted by this paper is not based on aggregate values such as mainstream current weighing text analysis but based on the distributed representation of words calculated by machine learning (fasttext). Therefore, the basis of the analysis is not the sum of the words in the text but the similarity using the distributed

representation of the word. Text mining based on distributed representation is still an incomplete technology and it is not a fixed method, it is a field where further development is expected in the future. In this paper, we propose two analytical methods using this distributed representation.

The method proposed by this paper is useful for business researchers, also for practitioners, when planning and model change of products/services, comparison with other company's product services becomes easier than ever, more detailed analysis.