様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　元 年　 ５ 月 １７ 日現在

機関番号：２４４０２
研究種目：若手研究(B)
研究期間：2017～2018
課題番号：１７Ｋ１３７８７

研究課題名（和文）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

研究課題名（英文）Visualization by text mining : Semantic Attributes Preferred in Innovation Diffusion

研究代表者
竹岡　志朗（TAKEOKA, Shiro）

大阪市立大学・大学院経営学研究科・経営学研究科付属先端教育センター特別研究員

研究者番号：７０７１１５５５

交付決定額（研究期間全体）：（直接経費）　　2,400,000円

研究成果の概要（和文）：　本研究は商品カテゴリーにおける諸属性の有無や高低が、個々の商品の魅力、つまり競争優位性にどのように影響しているのかを明らかにすることを目的とした。
　これを明らかにするために、本研究ではテキストマイニング、特に近年世界的に注目を集めているニューラルネットワーク技術に基づく分散表現テキストマイニングを用い、分析対象としてはインターネット上にあるクチコミデータを用いた。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ていることで、これまでの外形的評価基準とは異なる、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た。

研究成果の学術的意義や社会的意義
　これまでも競合間の関係などを可視化するためにテキストマイニングは用いられてきた。しかし、それは単語の出現回数や共起関係の強さの集計が基本でありそのような分析だけでは、その単語に付与された意味の分析はできなかった。本研究ではニューラルネットワーク基盤技術とする分散表現をテキストマイニングを使用することで、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った。
　また、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ていることで、これまでの外形的評価基準とは異なる、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た。

研究成果の概要（英文）：　The purpose of this research is to clarify how the presence or absence of various attributes in the product category and the level of the product attribute affect the attractiveness of the individual product, that is, the competitive advantage.
　In this research, we used text mining, in particular, distributed representation text mining based on neural network technology that has recently attracted worldwide attention, and used eWoM data on the Internet as an analysis target.
　By using words based on the experience of the consumer's eWoM, it is used based on the internal evaluation criteria created by the consumer's psychological experience different from the previous external evaluation criteria. The advantage is that we can use linguistic expressions for analysis.

研究分野：経営学

キーワード：テキストマイニング　普及　機械学習

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９、ＣＫ－１９（共通）

１．研究開始当初の背景

イノベーションは企業の持続的な競争優位の源泉である。イノベーションに関する研究は、大別するとベストプラクティス探求研究と普及研究に分類できる。ベストプラクティス探求研究は、新しいモノを生み出す組織の能力や、行動、文化、商業的に成功する方法などを明らかにすることを目的としている。他方、本研究も含まれる普及研究は、ミクロな行為（個人の知覚や行為）の集合として起こるマクロな社会の変化（普及という現象や普及率の上昇）を対象とし、S字カーブ(Bass, 1969)などの形で視覚化されることが多い。商品カテゴリーのライフサイクルを通じた伝播・浸透の過程を対象とし、その過程でどのようなことが起こり、どのように進んでいくのか、また、どのような要因がその成否や速度に影響を与えているのかを明らかにすることに主眼を置いている。普及研究が注目される理由としては、イノベーションは、その先行者が必ずしも利益を得るわけではなく、2 番手以降の模倣者がより利益を得ることも多く(Teece, 1986)、また、その普及過程でより経済的収益が高いのは、初期よりも漸進的に進歩する過程であるためである。

しかし、これまでの普及研究では、イノベーションの定義にそれに関与する消費者やメーカーが新しいと知覚するモノをおき、その伝播・浸透の過程を明らかにしようとしてきたにもかかわらず、分析の中で登場する消費者はその新しさや、優位性などの属性だけを知覚する消費者、あるいは採用・非採用だけを決定する消費者であった。言い換えれば、抽象化された、イノベーションに対する知覚や認知をかなりの程度制限された消費者であった。

２．研究の目的

本研究では、イノベーションの普及過程にアプローチするにあたって消費者の商品に対するクチコミを用いる。これによって、これまでの普及研究とは異なる、イノベーションを選択し、受容する消費者の認知、つまり購入し、使用する中で商品に対して付与する意味とその発露としてあらわれるテキストから、普及過程とその特徴を明らかにすることができる。

このような目的を実現するために、本研究では現在世界的にも注目を集めているニューラルネットワーク技術を用いたテキストマイニングを研究の手法として採用した。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 に関する言葉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採用した。今回採用した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。

３．研究の方法

今回採用した手法は、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、本報告書では分散表現テキストマイニングと呼ぶこととする。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。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)。

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。

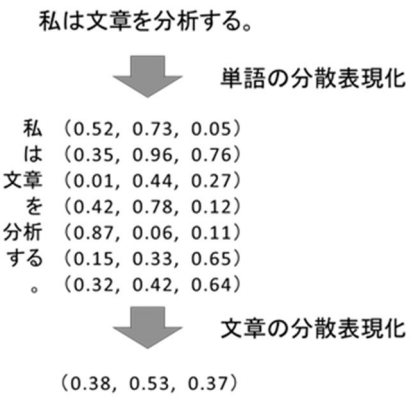

図 1　単語と文章の分散表現化

竹岡（2019, p123）

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ていることで、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。

しかし、類似度の測定や、単純な「類推問題(Mikolov et al, 2013)」を解く だけではテキストマイニングには不十分である。そこで、本研究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、施設の特徴を明確に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。

たとえば、「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、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。

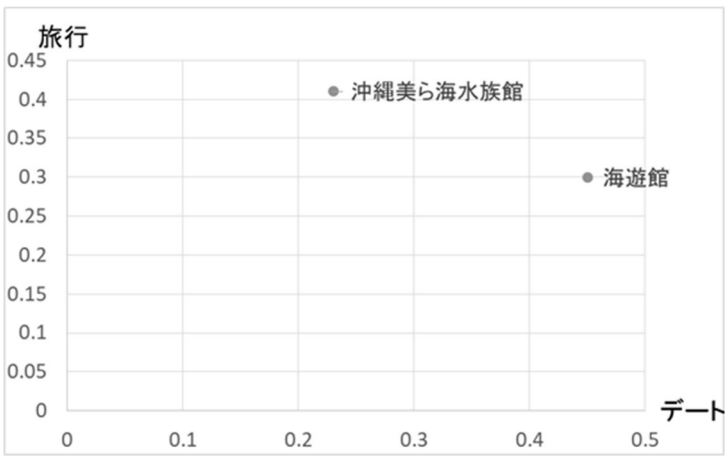

図 2 代理変数を用いた各施設の特徴の可視化

竹岡（2019, p126）

本研究では　本研究は商品カテゴリー（テーマパークや水族館のような施設）における諸属性の有無や高低が、個々の商品の魅力、つまり競争優位性にどのように影響しているのかを明らかにすることを目的としている。つまり、商品カテゴリーを構成する諸属性を意味的属性、つまり定量的に把握可能な外形的属性ではなく、消費者が使用の中で形成する意味的側面という観点からとらえ、またその諸属性が競争力に与える影響を明らかにしようとする点にある。そこで、分析対象としては水族館 5 施設、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

古屋港水族館を選択、実際の分析には「じゃらん net」に投稿されているものを用いた。
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 」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、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 。

４．研究成果

本研究は上記目的を達成するために 4 つのステップを踏んだ。第一に、インターネット上のクチコミを利用するにあたって、どの範囲のデータを集める必要があるのかを検討した。たとえば今回利用した「じゃらん net」のデータはインターネット上に投稿されている各水族館に関するクチコミのごく一部である。本来であれば、全クチコミを収集し、それを分析する必要がある。そこで、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』 Vol.69-1，pp.91-107.ではその可能性について検討した。その結果、4 つの点、つまり①キーワードを含む web ページを検索エンジンだけを頼りに見つけることは困難、②www 全体をクローリングできればキーワードを含む web ページを取り出し、それらをコミュニティとして定義することもできるが、困難、③キーワードの存在する web ページからたどることのできる web ページだけを取得するように設定することでの web コミュニティの析出は困難、④外部ドメインへの web リンクが少数しか存在せず、そのためクローリングによる web コミュニティの析出は困難、という点を踏まえて、本研究では「じゃらん net」に投稿されたデータだけを使用することとした。
続いて、第 2 ステップとして竹岡志朗（2018）「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 回大会』および高木修一、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』 Vol.64-2，pp.241-260.の中で用いたいような、分散表現を用いたテキストマイニングについて検討した。分散表現を用いたテキストマイニングの基本は単語間の類似度を用いた特徴の分析である。これにより、商品カテゴリー内の分析対象、つまり商品や施設などの特徴を明らかにすることが可能であることが分かった。
第 3 ステップは竹岡志朗（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-」『経済経営論集』 Vol59-4， pp.101-122.である。上記駄 2 ステップの方法では各施設や商品の特徴は明らかになる一方で、相対的な競争関係については十分に明らかにすることができなかった。そこで、この第 3 ステップでは代理変数として分散表現を用いる手法を開発し、これによって各施設の相対的特徴を知覚マップの形で可視化することに成功した。

表 1　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

最後に、第 4 ステップとして、竹岡志朗（2019）「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」『経済経営論集』 Vol60-4， pp.121-143.を行ったこの中では、上述方法にあるような代理変数としての分散表現を用いた分析手法に加えて、外形的データ、すなわち施設への入場者数などを併用することで施設間の競争関係と関連のある概念を析出する方法を開発し、競争関係の可視化に成功した。以下がその成果である。
外形的データとしては Wikipedia に掲載されている来場者数を利用した。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)、それらと上記 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 1 は相関の強い 20 概念を抽出したものである 。

上記分析結果から次のことがいえる。

- 「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）

以上が本研究の成果である。分散表現テキストマイニングを代理変数を用いて使用することで、これまでとは異なる分析が可能なる。特に、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。

５．主な発表論文等

〔雑誌論文〕（計 5 件）

1. 竹岡志朗（2019）「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」『経済経営論集』 Vol60-4, pp.121-143.
2. 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』Vol.64-2, pp.241-260.
3. 竹岡志朗、神谷栄司、土井捷三「量的観点から見る日本におけるヴィゴツキー研究の発展 ： 理論のイノベーションと普及」『ヴィゴツキー学』Vol.5, pp.1-13.
4. 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-」『経済経営論集』 Vol59-4, pp.101-122.
5. 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』 Vol.69-1, pp.91-107.

〔学会発表〕（計 4 件）

1. 竹岡志朗（2018）「機械学習を活用したテキストマイニング―概念間の相関分析による特徴の確認」『日本情報経営学会第 77 回大会』
2. 竹岡志朗（2018）「機械学習を活用したテキストマイニング―外形的データを併用することによる特徴分析」『日本経営学会第 92 回大会』
3. 竹岡志朗（2018）「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 回大会』
4. 竹岡志朗、高木修一（2018）「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』

６．研究協力者

研究協力者氏名：高木　修一

ローマ字氏名：TAKAGI, Shuichi