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　 ６ 月　 ５ 日現在

機関番号：２２６０４
研究種目：基盤研究(C)（一般）
研究期間：2016～2018
課題番号：１６Ｋ００００５２

研究課題名（和文）大規模複雑データの特性に即した解析法の選択と新解析法についての総合的研究

研究課題名（英文）Study of Selection of Model and Proposal of New Method Considering Attribute of Large Complex Data

研究代表者
中山　厚穂（Nakayama, Atsuho）

首都大学東京・経営学研究科・准教授

研究者番号：６０４３４１９８

交付決定額（研究期間全体）：（直接経費）　　3,600,000円

研究成果の概要（和文）：本研究課題では，複数の情報が組み合わされた構造や形式が複雑で大規模なデータの特性に適した解析のための方法論の構築と適切な解析法の選択法の提案を目指して研究を実施した．最近、多くの分野で多種多様な情報を含む大規模で複雑なデータの分析への関心が高まっている。大規模で複雑なデータの属性を適切に分析するための方法の提案を行った。そして、多くの複雑な社会現象や心理現象を扱う学問分野について提案方法を応用した．本研究の成果をもとに、大規模複雑なデータを分析することのニーズが高まっている多くの分野での貢献を目指した．

研究成果の学術的意義や社会的意義
多相多元データをより低次の相や元に圧縮・分割した上で分析すること，非対称データにおいて非対称モデルで分析することの是非を判断するための方法論の構築により構造や形式がより複雑なデータを解析可能なモデルを使うべきなのか，単純なモデルで解析を行うべきなのかということについての指針を示すことができた．大規模複雑なデータを分析することのニーズが高まっている多くの分野での貢献が期待される．本研究における大規模複雑なデータを解析するための方法を提案することがき，多くの複雑な社会現象や心理現象を扱う学問分野あるいは応用分野への適応が期待できる．

研究成果の概要（英文）：In this study, we aim to study the framework of the selection of the appropriate statistical model to analyze large complex data considering the attribute of the structure and form of the data. Recently, there has been increasing interest in the analysis of large complex data that contain a wide variety of information in many fields. We proposed the method that can appropriately analyze the attribute of large complex data. We applied the proposed method to the academic field dealing with many complex social and psychological phenomena. Based on the results of this study, we tried to contribute in various fields where the need to analyze large complex data is increasing.

研究分野：統計科学

キーワード：大規模データ　　多相多元データ　スパースデータ　　非対称データ　次元縮約　　クラスタリング　多次元尺度構成法

様　式　Ｃ－１９，Ｆ－１９－１，Ｚ－１９，ＣＫ－１９（共通）

１．研究開始当初の背景

社会現象や心理現象を扱う様々な学問分野あるいは応用分野で，調査項目，変数，回答者，データ収集時点など複数の情報が組み合わされた構造や形式が複雑で大規模なデータが入手可能となっている．これらの大規模複雑データをどのように活用していくのか，どのように分析するのかということが喫緊の課題となっている．データの構造や形式は，データの相と元の数によって分類することができる．相は，一組の対象(調査項目，変数，回答者，データ収集時点など)を意味し，相の数は，データに相がいくつ(何種類)存在しているのかを示す．元の数は，相がデータで合計何個組み合わされているのかを表す．また，データの対角に沿って対応する要素が等しいデータは対称データ，異なるデータは非対称データとなる．データを分析する際には，各データの構造や形式に適した手法で分析することが望ましい．しかし，実際には，例えば対象×個人×時点からなる 3 相 3 元データを分析する際に，1 つの元の圧縮や何らかの変換により 2 相 2 元データとした上で分析するといったような，データの形式に適さない方法が用いられる場合が多く存在する．さらに，このような圧縮や分割の是非を吟味してから分析が行われている場合もあるが多くの場合には，このような吟味は行われていない．多相多元データの元(もしくは相)を圧縮することは，その元(相)がもつ情報を利用しないこと，あるいは，それ以外の元(相)が表すデータが圧縮された元(相)について同一，つまり，圧縮された元(相)における変動は誤差であるとみなすことになる．また，2 元データなどに分割し，各々のデータを個別に分析すると，各データの分析結果が独立に得られ，個々の分析結果には関連がなく，結果の解釈が難しいことが多い．多相多元データのようなデータの構造や形式が複雑なデータを分析する際には，データの構造や形式に適した手法を用いて，圧縮や分割をせずに分析することが望ましい．このことは，非対称データをそのまま分析せずに対称化して分析する場合においても共通した問題である．一方で，多相多元データの構造や形式に適した手法によりすべての情報を反映した解析を行うことができたとしても，分析で得られた情報をすべて正しく理解できるという保証はなく，社会現象や心理現象をより単純なシンプルなモデルで表現をした方が分析結果を一般化しやすい可能性も高い．また，多相多元データにおいてデータの規模が大きい場合にはデータがスパースとなることも想定され，計算時間の短縮や分析結果の効率的な解釈方法についても考慮する必要がある．したがって，大規模複雑データの構造や形式に即した解析法を様々な解析手法の中から取捨選択する必要がある．

これまでに行ってきた Nakayama(2013)・Nakayama (2015)などの研究では，高次のモデルとより低次のモデルによる分析結果を比較検討することや各元の交互作用を考慮することで，データを圧縮・分割することが妥当かを検証するための方法論を提案し，その妥当性を示した．しかし，Nakayama(2013)や Nakayama(2015)では，多元データの分析とその分析の妥当性の評価についての分析は個別に行われている．各元･各相間の交互作用をより正確にとらえるためには，それぞれの分析を個別に行うのではなく同時に行うことのできる方法論の提案が必須といえる．また，Nakayama, and Okada(2012a)や Nakayama, Tsurumi, and Okada(2013)では単相 3 元非対称データの構造や形式に即した解析法の選択のための指針を導くための方法論を提案し，その有効性を示した．しかし，Nakayama, and Okada(2012a)や Nakayama, Tsurumi, and Okada(2013)はデータの構造や形式に即した解析法の選択のための方法論としての有効性を示しているが，単相 3 元データについての限定的な研究であり，より高次のデータへの拡張が望まれる．そして，中山他(2008)，鶴見・中山・増田(2013)，中山・増田・鶴見(2015)ではソーシャルメディア上の書き込みから作成した文章×単語からなる大規模スパースなデータに対してトピックモデルを適応し，書き込み内の話題の抽出や時系列的な変化の把握を試みている．しかし，大規模スパースなデータをトピックモデルにより次元縮約を行うことで解析する場合において，分析に用いる対象や変数の数が膨大となると,結果の解釈が困難になるという問題も存在する．得られた分析結果をより効果的に表現するための方法の提案が必要といえる．

２．研究の目的

本研究課題では，複数の情報が組み合わされた構造や形式が複雑で大規模なデータの特性に適した解析のための方法論の構築と適切な解析法の選択法の提案が目的である．様々な分野で，大規模複雑データを分析することへの需要が高まるとともに，データの特性に即した解析を実行することが求められている．これらのことを実現するための研究が行われているが，既存研究で扱われている手法が一般の応用分野に普及しているとはいいがたく，どのような特性のデータにはどのような解析法を適用すべきなのかという指針は確立されていない．以上から，本研究課題では「多相多元データ」,「非対称データ」,「スパースデータ」の構造や形式に即した解析法の選択のための方法論と新解析法の提案を目的として研究を行う．

３．研究の方法

本研究課題では，大規模複雑データ，特に「多相多元データ」,「非対称データ」,「スパースデータ」の構造や形式に即した解析法の選択のための方法論の構築と新解析法の提案を目指した研究を実施した．まず，各データの有している特性について整理し，その特性の理論的特徴付けを行った．そして，既存の解析法と各データが有している特性との対応付けを実施した．その成果を踏まえ，各データの構造や形式に即した解析法の提案と各データの特性に適した解析法を選択するための方法論の確立を目指した．また，提案した方法論の数理的妥当性につい

て検討をするとともに，様々な実データに適応することで応用分野への適応可能性について検証した.

４．研究成果

大規模複雑データを分析する際の適切な手法の選択法についての研究では，大規模データのためのクラスタ中心を再計算しない非階層的クラスタリング法を提案することによって，大規模データの特性に合わせてクラスタリングを効率的に行う方法についての研究を行った．中山・出口・烏谷 (2016)， Nakayama and Deguchi (2017)において大規模データのためのクラスタ中心を再計算しない非階層的クラスタリング法を提案することにより，大規模データの特性に合わせてクラスタリングを効率的に行う方法についての研究を行った.

大規模な多相多元データを分析する際にはデータがスパースとなり，データのもつ情報を反映した分析がスムーズに実施できない可能性も考えられ，この課題を解決するための研究を実施した. Nakayama (2016), 中山（2016）, Nakayama (2017), Nakayama (2018a), Nakayama (2018b), Nakayama, Paliwoda, and Smolak (2019) では Twitter の書き込みデータを対象として研究を実施した．これらの研究を通じて大規模複雑データから有益な知識を抽出するためなの解析法の構築や提案を行った．また，Nakayam and Baier (2018b)や Nakayam and Baier (2019)では Web 上の画像や動画データを対象としてニューラルネットワークを用いた研究を行った. 研究を行った．その成果をもとにブランド間のイメージの関連性の解明のための応用を目指した．そして，多相多元データの特性を考慮した研究については, Nakayam and Baier (2018a)において Three-mode Data を分析するための Two-mode Overlapping Clustering についての研究を行った．得られた知見をもとに，オンラインショッピングのサイト改善のための提言を行った.

多相多元データを分析する際に，多相多元データの圧縮・分割すべきかどうかの是非についての研究を Nakayama (2016)において, データの特性に即した適切な分析手法の選択についての方法を提案した．また，非対称データの分析の際の適切な手法の選択法についての研究も行った中山 (2016)において実施し，データの特性に即した手法による分析方法についての提言を行った.

５．主な発表論文等

〔雑誌論文〕（計　2 件）

1．Nakayama, A. (2017). The Classification and Visualization of Twitter Trending Topics Considering Time Series Variation. In F. Palumbo, A. Montanari, M. Vichi, (Eds.), Data Science - Innovative Developments in Data Analysis and Clustering (pp. 161-173). Heidelberg-Berlin, Springer. DOI：10.1007/978-3-319-55723-6, 査読有

2．Nakayama, A. (2016). Evaluating the Necessity of a Triadic Distance Model. In A. F. X. Wilhelm, and H. A. Kestler (Eds.), Analysis of Large and Complex Data (pp. 195-205). Heidelberg-Berlin, Springer-Verlag. DOI：10.1007/978-3-319-25226-1, 査読有

〔学会発表〕（計　14 件）

1．Nakayama, A., Paliwoda-Matiolańska, A., and Smolak-Lozano, E. (2019). The text mining and dimension reduction method application into exploring the isomorphic pressures in the corporate communications on the textual tweet data about sustainability in the energy sector [summary]. Abstracts of European Conference on Data Analysis (ECDA2019), p. 52 (March 20, 2018).

2．Nakayama, A., and Baier, D. (2019). Beer brand image classification using deep learning [summary]. Abstracts of European Conference on Data Analysis (ECDA2019), p. 54-55 (March 19, 2018).

3．Nakayama, A. (2018a). The classification and visualization of trending topics in online word-of-mouth data [summary]. Abstracts of the 23rd International Conference on Computational Statistics (COMPSTAT2018), p. 29 (August 30, 2018).

4．Nakayama, A. (2018b). Topic detection and classification in consumer web communication data [summary]. Abstracts of European Conference on Data Analysis (ECDA2018), p. 88 (July 4, 2018).

5．Nakayama, A., and Baier, D. (2018a). Two-mode overlapping clustering for three-mode data with applications to online shopping and site engineering [summary]. Abstracts of European Conference on Data Analysis (ECDA2018), p. 75 (July 4, 2018).

6. Nakayama, A., and Baier, D. (2018b). Image data analysis in web communication data using deep learning [summary]. Abstracts of German-Japanese Symposium (July 3, 2018).

7. 菰田文男・中山厚穂 (2017). ピンポイントフォーカス型テキストマイニング手法の研究, 第 11 回 テキストアナリティクス・シンポジウム (2017 年 9 月 7 日).

8. 中山厚穂・増田純也・鶴見博之 (2017). Web コミュニケーション・データ上のトピックの分類と変化についての分析,日本行動計量学会第 45 回大会(2017 年 8 月 30 日).

9. Nakayama, A. and Komoda, F. (2017). Study on Globalization of Japanese Companies Based on Text Mining [summary]. Abstracts of 6th Japanese-German Symposium on Classification, p. 12-13, (August 12, 2017).

10. Nakayama, A. and Deguchi, S. (2017). Non-Hierarchical Clustering for Large Data without Recalculating Cluster Center [summary]. Abstracts of the 2017 Conference of the International Federation of Classification Societies, p. 198 (August 8, 2017).

11. Nakayama, A. (2016). Analysis of Trending Topics in Consumer Web Communication Data [summary]. Abstracts of German-Japanese Symposium 2016. (September, 12, 2016, Invited session).

12. 中山厚穂 (2016). マーケティングにおける Web コミュニケーションデータ活用の可能性, 2016 年度 統計関連学会連合大会, 2016 年 9 月 5 日.

13. 中山厚穂 (2016). マーケティングデータにおける 非対称性の分析-Web 上のマーケティング・コミュニケーションデータの分析-, 日本行動計量学会第 44 回大会, 2016 年 8 月 31 日.

14. 中山厚穂・出口慎二・烏谷正彦 (2016). クラスタ中心を再計算しない 大規模データのための 非階層的クラスタリング,日本行動計量学会第 44 回大会, 2016 年 8 月 31 日.

〔図書〕（計　0 件）

〔産業財産権〕
○出願状況（計　0 件）

名称：
発明者：
権利者：
種類：
番号：
出願年：
国内外の別：

○取得状況（計　0 件）

名称：
発明者：
権利者：
種類：
番号：
取得年：
国内外の別：

〔その他〕
ホームページ等

６．研究組織

(1)研究分担者
研究分担者氏名：
ローマ字氏名：

所属研究機関名：

部局名：

職名：

研究者番号（8 桁）：

(2)**研究協力者**

研究協力者氏名：鶴見裕之

ローマ字氏名：(Hiroyuki, Tsurumi)