科学研究費助成事業　　研究成果報告書

平成 ２６ 年　 ５ 月 ２９ 日現在

機関番号：１３９０１
研究種目：若手研究(A)
研究期間：2011～2013
課題番号：２３６８００２６

研究課題名（和文）ソーシャルメディア発信情報のバイアス補正法の研究

研究課題名（英文）Research on bias adjustment methods of data from social media

研究代表者
星野　崇宏（HOSHINO, Takahiro）

名古屋大学・経済学研究科（研究院）・准教授

研究者番号：２０３９０５８６

交付決定額（研究期間全体）：（直接経費）　　10,200,000円、（間接経費）　　3,060,000円

研究成果の概要（和文）：本研究では、ソーシャルメディア上における書き込み情報や、blogのトラックバックなどを含めたWeb上のソーシャルネットワーク情報の偏りを補正するための基礎的な方法論を開発し、そのためのデータ収集方法と解析方法を定式化することを目的とする。具体的にはソーシャルメディアの発信者が「本来対象とする母集団」から偏って抽出されている”選択バイアス”が生じているという問題定式化のもとに、統計学や計量経済学で近年解析方法が盛んに研究されている選択バイアスモデリングの考え方を利用した手法の開発を行った。研究成果の一部は統計学のトップジャーナルに掲載された。

研究成果の概要（英文）：In this research we planned to develop fundamental methods to reduce biases inherent in data from social media such as opinion data in consumer generated media. To be more concrete, we developed some statistical methods to reduce such biases by regarding them to be the sample selection bias in that the originators of the texts in consumer generated media are nonrandomly sampled from the targeted population the analysts of the data would like to know about. One of the proposed methods is published in the top journal in statistics (Journal of the American Statistical Association).

研究分野：統計科学

科研費の分科・細目：図書館情報学・人文社会情報学

キーワード：選択バイアス　因果効果　傾向スコア　マーケティングリサーチ

１．研究開始当初の背景

Web 上に存在する Blog やソーシャルネットワーキングサービス、電子掲示版などのソーシャルメディア（あるいは消費者生成メディア）は近年では人々の消費活動や日常の余暇活動、さらには投票に至るまで、意思決定に影響を大きく与える非常に重要な情報源となりつつある。実際、消費者は Web 上のネガティブな評判をもとに、特定の製品やサービスの購入を差し控える行動をとることが知られている。一方でポジティブな評判についてはそれをほとんど信じず、ソーシャルメディアでの評判情報は購買意欲を減退させる効果が大きいことが様々な実証研究によって示されている(例えば濱岡・里村,09)。さらにマーケティング、政治学、社会心理学、経営学などの研究者、さらには政府機関までもがネット上でのソーシャルメディアから得られる情報の利活用を検討しつつある。

ソーシャルメディアから得られる情報量は量的には非常に膨大であるが、そもそもこのようなネット上の情報発信は「書きたい人だけが書く、答えたい人だけが答える」といった形のものである。本来関心のある消費者や国民全体の意見や購入頻度などを知りたいという目的からは質の高い情報ではなく、情報の信頼性が大きく問われている。実際、Blog の書き込みを行っている人の年齢や職業、年収などの分布は一般消費者から大きく乖離しており（例えば星野,2008）、政府・自治体の公共マーケティングや企業の新製品開発のための情報源として信頼されるには至っていない。

ソーシャルメディアにおける発信情報そのものの信頼性評価については、発信者の属性推定、クラスタリングの研究が行われている（たとえば奥村,08；小山,09；Bhattacharya,06;Bollegala,08）が、これまでの中心的な関心は Web ページデータベースにおけるレコードリンケージを中心としたアイデンティティ特定であった(相澤,05;Winkler,06)。また、blog サイトや SNS サイト、EC サイトなどでの登録 ID を複数の組織間（ドメイン間）で共有する ID 連携などもアイデンティティを特定することなく同一個人の行動や情報発信をリンクさせる方法として研究されている(下江,09 など)が、個人情報保護法上の問題や利用者の拒否感、加えてたとえ ID 連携をしても参加者の偏りは残るといった問題がある。一方、Web ハイパーリンク構造分析に基づいて「信頼される Web サイト」へのリンクがあるサイトの情報は信頼できる（＝信頼性伝搬法）、という議論があるが、ソーシャルメディアの発信情報では「信頼される Web サイトおよび発信者」を同定するのは困難である。ある製品やサービスが望ましい／購入したい、あるいはある政党を支持する、という情報発信において「信頼される代表的なサイトや発信者」を定義することは不可能であるからである。関連研究は情報通信分野の研究者のみが行っているため、「ソーシャルメディアの発信者」が世論理解、景況判断や新製品開発などにおける「母集団」に対して極めて偏った標本である、いう発想が欠如しており、この問題はこの数年の間に理解されるようになったばかりである。

２．研究の目的

本研究では、ソーシャルメディア上における書き込み情報や、blog のトラックバックなどを含めた Web 上のソーシャルネットワーク情報の偏りを補正するための基礎的な方法論を開発し、そのためのデータ収集方法と解析方法を定式化することを目的とする。具体的にはソーシャルメディアの発信者が「本来対象とする母集団」から偏って抽出されている”選択バイアス”が生じているという問題定式化のもとに、統計学や計量経済学で近年解析方法が盛んに研究されている選択バイアスモデリングの考え方を利用した手法の開発を行うことを目的とする。

特に、選択バイアスを欠測データとみなす近年の統計的欠測データ解析の考え方を適用するが、書き込む側が特定の意図を伝えたいという意思が強いほど書き込まれるということが想定されるため、書き込み有無を欠測データとみなす場合にはランダムでない欠測を考慮した解析法の開発が必須であることがわかった。そこでモデル仮定の少ない

３．研究の方法

ソーシャルメディアの発信情報の偏りを補正するためには、ソーシャルメディアの発信者が「本来対象とする母集団」から偏って抽出されている“選択バイアス”が生じていると考え、統計学・計量経済学を中心に近年解析手法の盛んに研究されている選択バイアスモデリングの考え方をもとに、研究目的に従った選択バイアスモデリングに基づく補正法の開発と実データ、およびシミュレーションによる妥当性検証を行う。具体的には様々な共変量情報を用いて選択バイアスを補正する方法として傾向スコア解析の有用性を調べる。但し書き込む側が特定の意図を伝えたいという意思が強いほど書き込まれるということが想定されるため、書き込み有無を欠測データとみなす場合にはランダムでない欠測を考慮した解析法の開発が必須である。そこでセミパラメトリックベイズモデリングを用いた選択バイアスモデルを開発する。

さらにソーシャルメディアでの発信情報及びネットワーク情報についての取得、発信者の様々な情報（共変量と呼ぶ）の取得、偏りのない抽出法によって得られた対象者集団に対する共変量情報の取得、少ない共変量

情報で精度よく補正を行うための共変量選択法の開発、実データを用いた補正の再現性の検証研究、実務に応用可能な、選択バイアスモデリング手法と共変量選択法の近似的手法の開発を行う。

４．研究成果

ネット調査やウェブの書き込み情報について、ネット調査参加や書き込み時や予測される属性などの様々な共変量情報を用いて選択バイアスを補正する方法として傾向スコア解析があるが、その有用性を調べたところ、傾向スコアを用いた補正だけでは偏りは十分除去できない場合があることがわかった。これは回答する側や書き込む側が特定の意図を伝えたいという意思が強いほど回答する、書き込まれるということであり、書き込まれるであろう内容自体によってその書き込みが実際に投稿されるという点で、書き込みの欠測がランダムでない欠測であるということを示唆している。そこで我々はランダムでない欠測を考慮した解析法の開発が必須であると考え、セミパラメトリックベイズモデリングを用いた選択バイアスモデルを開発した。提案したモデルが先行研究と異なる点は、共変量 x そのものが固定値ではなく確率変数とみなし､目的となる変数 y と z=1 ならば観測、z=0 ならば欠測とする欠測インディケータ変数 z の同時分布を

$$p(y,z,x) = p(z \mid y,x)p(x \mid y)p(y)$$

のように分解し、さらに上記右辺の第一、第二項のモデル仮定が難しいことからこの部分については分布仮定を置かないノンパラメトリックベイズ法を新しく開発したという点である。この結果を実際にインターネット閲覧情報データに適用し、口コミの効果についてこれまでも先行研究で言われているように同類効果が大きいこと、そしてその効果を除去した場合に口コミの真の効果は小さいことを示した。

さらにソーシャルメディアでの発信情報（意見記載した、行動記述）及びネットワーク情報についての取得を行い、そこから発信者の様々な情報（共変量と呼ぶ）を取得し、少ない共変量情報で精度よく補正を行うための共変量選択法の開発を行った。これらの研究の一部は現在投稿中である。

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計　１０件）

Takehiro Nagai, Takahiro Hoshino and Keiji Uchikawa (2011).
"Statistical Significance Testing with Mahalanobis Distance for Thresholds Estimated from Constant Stimuli Method".
***Seeing and Perceiving,* 24,** 91-124 .

Shiro Ojima, Hiroko Matsuba-Kurita, Naoko Nakamura, Takahiro Hoshino and Hiroko Hagiwara (2011).
"Age and amount of exposure to a foreign language during childhood: Behavioral and ERP data on the semantic comprehension of spoken English by Japanese children".
***Neuroscience Research,* 70,** 197-205.

Tetsuro Kobayashi and Takahiro Hoshino (2011).
"Propensity Score Adjustment for Internet Panel Surveys of Voting Behavior: A Case in Japan".
***Japanese Journal of Electoral Studies.* 27(2)** 104-117.

猪狩良介･星野崇宏 (2012)
“非集計 Web アクセスデータを用いたサイト普及モデル：多時点・複数サイトの階層ベイズモデリング”
マーケティング・サイエンス, 第 20 巻 1 号, 43-67.

Takahiro Hoshino. and Peter M Bentler (2013). “Bias in Factor Score Regression and a Simple Solution".
In ***Analysis of Mixed Data : Methods & Applications*** *(A.R. de Leon & K. C. C. Carriere,eds)*. NY: CRC Press, 43-61.

宮崎慧･星野崇宏 (2013)
“複数商品購買行動理解のための階層

ベイズグレンジャー因果性分析”
マーケティング・サイエンス, 21, 11-35.

Takahiro Tabuchi, Takahiro Hoshino , Tomio Nakayama, Yuri Ito, Akiko Ioka, Isao Miyashiro, and and Hideaki Tsukuma (2013).
“Does removal of out-of-pocket costs for cervical and breast cancer screening work? A quasi-experimental study to evaluate the impact on attendance, attendance inequality and average cost per uptake of a Japanese government intervention”
***International Journal of Cancer,* 133,** 972-983.

星野崇宏 (2013)
“継続時間と離散選択の同時分析のための変量効果モデルとその選択バイアス補正：Web ログデータからの潜在顧客への広告販促戦略立案”
日本統計学会誌, 43, 41-58.

Takahiro Hoshino. (2013).
“Semiparametric Bayesian Estimation for Marginal Parametric Potential Outcome Modeling: Application to Causal Inference"
***Journal of the American Statistical Association,* 108,** 1189-1204.

猪狩良介・星野崇宏 (2014)
“**階層ベイズ動的サンプル・セレクションモデルによる Web サイトへの誘導とサイト閲覧行動の同時分析**”
日本統計学会誌, 43, 185-214.

〔学会発表〕（計　9　件）

宮崎慧・星野崇宏(2011)
「動的階層ベイズモデルを用いた複数商品カテゴリー購買とブランド購買の同時分析」統計関連学会連合大会 2011 年大会、九州大学

太田悠大・星野崇宏 (2011)
「プロスペクト理論を考慮した同時購買行動での価格プロモーション戦略」
第 90 回日本マーケティング･サイエンス学会研究大会、電通ホール

Kei, Miyazaki and Takahiro Hoshino. (2011)
“Hierarchical Bayes Granger Causality Analysis for Understanding Purchase Behaviors”
The 76th Annual and the 17th International Meetings of the Psychometric Society, Hong-Kong Institute of Education, Hong-Kong.

Takahiro Hoshino. (2012)
“Causal Inference for Multilevel Modeling Under Nonrandom Allocation to Level-2 Units: Moderated causal effect as a function of macro-level variables”.
IMPS2012, The 77nd Annual and the 18th International Meetings of the Psychometric Society, Lincoln, NE. USA.

宮崎慧・星野崇宏(2012)
「ブランド非購買に対する家庭内在庫変数およびブランドスイッチングの効果の分離と推定」
統計関連学会連合大会 2012 年大会、北海道大学

星野崇宏(2012)
「ポイントプログラムの長期効果：目標

勾配仮説は成立するのか」
　第６回行動経済学会大会、青山学院大学.

飯塚久哲・星野崇宏・鈴木重央・大黒未鈴(2013)
「データ融合を用いたシェア・オブ・ウォレットの推定」
　第 47 回消費者行動研究コンファレンス、法政大学

竹内真登・星野崇宏(2013)
「解釈レベルの操作を伴うマーケティングリサーチ手法の開発とバイアスの排除に関する実証実験」 第 47 回消費者行動研究コンファレンス、法政大学

宮崎慧・星野崇宏(2013)
「商品カテゴリー購買と複数ブランド購買の段階型同時分析モデルの提案」
第 94 回マーケティングサイエンス学会、電通ホール

〔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：
出願年月日：
国内外の別：

○取得状況（計　０　件）

名称：
発明者：
権利者：
種類：
番号：
取得年月日：
国内外の別：

〔その他〕
ホームページ等

６．研究組織
(1)研究代表者
星野崇宏（HOSHINO　Takahiro）
　名古屋大学・大学院経済学研究科准教授
　研究者番号：20390586

(2)研究分担者
（　　）

　研究者番号：

(3)連携研究者
（　　）

　研究者番号：