消費者間コミュニケーション構造に関する計算社会科学的研究

［継続研究］

常勤研究者の部

代表研究者　水　野　　誠
明治大学
商学部
教授

共同研究者　瀧　川　裕　貴
東京大学大学院
人文社会系研究科・文学部
准教授

1章　はじめに

本研究は、ソーシャルメディア上で観察される消費者あるいは生活者のコミュニケーションを分析することで、その背後にある社会構造とりわけ社会階層の影響を解明したいという問題意識に基づいている。そのため、従来の社会科学や社会学における研究を踏まえつつ、計算社会科学（Computational Social Science）と呼ばれるアプローチを導入する。それは、人間行動に関して集積された膨大なデジタル情報（デジタルトレース）あるいは「ビッグデータ」を高度な計算モデルによって分析し、社会科学的な知見を獲得する文理横断的な研究の動きである。この手法を社会階層の解明に適用した研究は、われわれの知る限り皆無に近い。

このような研究を行う背景には、現代社会においてソーシャ

ルメディアの影響力がますます高まっていることがある。１つには、フェイクニュースや誤情報の蔓延、特定の国による情報操作といった政治的影響力の問題だが、もう１つは文化的影響力で、人々の感情やアイデンティティに与える影響が議論されている。われわれは、ソーシャルメディア上での文化的実践（発言、投稿）が、現実の社会における社会的差異（性や年齢、あるいは所得や富、職業などの社会経済的属性に基づく人々の差異）とどう対応しているかに注目する。ソーシャルメディア上で人々の社会的差異がシグナルとして発信されている可能性を踏まえて、Twitter への投稿（ツイート）からユーザーの個人属性をどの程度予測できるかを検証する。また、ツイートに現れるブランドへの言及と社会的差異の関係についても分析する。

本研究は先行研究のレビューから始まり、社会経済的属性のなかでも職業上の差異をより精細に捉えるべく、詳細な職業コードの設定と多次元のスコアリングを行う。次いでウェブ上でサーベイ調査を行って、個人属性と Twitter ユーザーのスクリーンネームを把握する。それを用いてツイートやプロファイルを収集し、個人属性の情報との関係を分析する。以上のフローを図示すると図１のようになる：

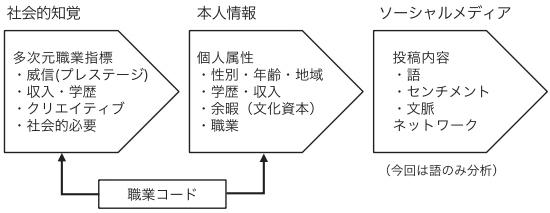

図１　本研究のフロー

2 章　先行研究のレビュー

社会的差異について扱う上で重要な先行研究は、社会学あるいは社会科学全般における社会階層研究である。古典的な研究としては Marx や Weber があるが、文化の実践や消費という観点が本格的に取り上げられたのは 20 世紀に、Bourdieu の研究が現れてからである。そこで重要な役割を担った文化資本の概念は、日本を含む様々な社会階層研究に継承されている。一方で、Bourdieu が経験的な分析の対象とした約 50 年前のフランスの状況と今日の文化状況は大きく異なることがすでに指摘されている。われわれの研究では、こうした研

究の流れを踏まえて、より今日的なコミュニケーションの状況に目を向ける。

研究の方法論として本研究が依拠するのが、冒頭でも述べた計算社会科学である。そこではソーシャルメディア、とりわけ Twitter から得られるデータが当初から重要な研究対象となってきた。しかし、ソーシャルメディア・データを用いて社会階層を分析した先行研究は少ない。その理由の１つは、分析上必要な個人属性がソーシャルメディアのデータに明示的に含まれていないことである。そこでいくつかの先行研究に倣い、ソーシャルメディア・データとサーベイ・データを組み合わせる方法を採用する。さらに疑似予測法と呼ばれるアプローチを適用し、ツイートから機械学習によって Twitter ユーザーの人口学的・社会経済的属性を予測する。予測に貢献した語は、当該ユーザーの社会的差異のシグナルになり得ると考える。

3 章　職業の多次元指標とその分析

社会階層の基礎にある職業を捉えるため、労働政策研究・研修機構が運営する「日本版 O-NET」から職業を抽出し、さらにいくつかの観点で職業を追加して 500 近い職業コードを設定した。それらに対して、ウェブ調査やクラウドソーシングでの作業を通じて、各職業に関する社会的知覚を多次元の指標で測定した。いずれも回答者本人の職業に対する主観的評価ではなく、ランダムに割り当てた職業に対する「客観的」評価を聴いている。

そうして測定した職業指標は、職業威信、権力、収入、学歴、クリエイティブネス、社会的必要性などである。職業威信は社会における階層的な地位を示す指標として、日本の社会階層研究において有名な SSM 調査などで用いられてきた。ここでは年収や学歴についても、その職業に相応しいと知覚される水準を聴いている。クリエイティブネスは Florida によるクリエイティブ・クラス論を踏まえ、知覚された水準を測定した。社会的必要性を指標に加えたのは、Graeber のブルシットジョブ論やコロナ禍におけるエッセンシャルワーカーへの評価を反映している。各職業について全回答者の平均を指標のスコアとした。その妥当性については、クリエイティブ／社会的必要性スコアについては日本版 O-NET が提供している職業の価値観や興味に関するスコアとの関係を分析して意味上の一致を確認している。クリエイティブ・スコアについては別の独立した調査でも測定したが、得られたスコアは高い相関を示した。

得られた職業ごとの多次元指標に主成分分析を行うと、それらは３次元に縮

約されることがわかった。主成分１は職業威信、権力、収入、学歴などと関係が深いので「階層」の次元と解釈した。同様に主成分２は「社会的必要性」、主成分３は「クリエイティブ」と解釈できる。

表１ 主成分１と２の上位 10 の職業

	主成分１（階層）		主成分２（社会的必要）	
順位	職業	スコア	職業	スコア
1	弁護士	5.87	救急救命士	3.87
2	裁判官	5.72	消防官	3.86
3	内科医	5.65	看護師	3.72
4	外科医	5.45	陸上自衛官	3.69
5	外交官	5.38	海上保安官	3.63
6	小児科医	5.08	警察官（都道府県警察）	3.58
7	会社経営者	5.05	海上自衛官	3.55
8	精神科医	4.99	訪問介護員/ホームヘルパー	3.04
9	宇宙開発技術者	4.85	外科医	2.92
10	パイロット	4.83	航空整備士	2.90

表２ 主成分３の上位 10 の職業

	主成分３				
順位	職業	スコア	順位	職業	スコア
1	和菓子製造、和菓子職人	3.19	6	理容師	2.60
2	洋菓子製造、パティシエ	3.08	7	酪農従事者	2.59
3	大工	3.02	8	パン製造、パン職人	2.55
4	日本料理調理人（板前）	2.90	9	イラストレーター	2.46
5	果樹栽培者	2.89	10	西洋料理調理人（コック）	2.45

各主成分スコアで上位にくる職業を挙げたのが表１と表２である。クリエイティブ・スコア自体は主成分１と弱いながらも正の相関があるので、クリエイティブな職業でも階層的に上位にある職業は、主成分３のスコアが必ずしも高

くならないことに注意したい。

4章　Twitterユーザーの特性と投稿の分析

4.1 Twitterユーザーの属性把握

Twitterで月２～３回以上投稿している20～69歳の男女(N=6,000)を対象にウェブ上でのサーベイ調査を実施した。主要課題の１つは、彼らを500近い職業コードのどれかに対応づけることである。今回はキーワード検索と多段階の多肢選択、そして自由回答を組み合わせた方法で本人職業を識別している。もう１つは、彼らのTwitterでのスクリーンネームを取得することである。

回答者の職業コードが識別できれば、３章で得た職業指標スコアや主成分スコアを当てはめることができる。職業指標の３つの主成分スコアを十分位数(デシル）に変換し、性別、年齢、年収（本人申告）との関係を見たところ、次のことがわかった。

- どの主成分でもデシルが上位になるほど女性比率が下がる（例外は「社会的必要」で、最上位のデシルでは女性比率が高まる）
- どの主成分も年齢とはほとんど関係しない
- 「階層」のデシルが高くなるほど年収は増える（他の主成分は年収と関係しない）

4.2 ツイートによる個人属性の予測

ウェブ調査で取得したスクリーンネームを用いて、Twitter REST APIから該当アカウントによって投稿された最大3,200のツイートを取得した。サーベイとの統合に成功した有効なアカウント数は3,930であった。さらに、各ユーザーのAPIで取得可能な限度内(最大3,200)での総ツイート数を調べたところ、20%近くのユーザーが100未満のツイート数であったため、これらのユーザーを分析から除外した。結果として、最終的に分析されたユーザー数は3,261となった。ツイート本文に、ストップワードの除去、形態素解析、最頻出語5000に限定、tf-idfによるウェイトづけといった前処理を行った。また、各ユーザーの投稿ツイートはひとつにまとめ１文書とした。

本研究で分析手法として採用した疑似予測法は、あるカテゴリをよく特徴づける特徴量の集合を探索的に発見するために予測の枠組みを用いる方法であ

る。例えば、議員の発言から、共和党議員か民主党議員かを予測することで、それぞれのイデオロギーを特徴づける語を発見するといった応用例がある。

疑似予測法は、形式としては教師あり機械学習の一種となる。ここで取り組む問題は、データの特徴量 X（この場合、ツイート本文やプロフィールのテキスト）から、目的変数 Y（この場合、ユーザーの人口学的、社会経済的属性）を予測することである。疑似予測法では、実際には、目的変数Yの値も既知であるが、あえてこれを未知のものとして予測を試みるのである。本研究における予測の対象 Y としては次の変数を選んだ。カテゴリ変数としては、「性別」「二値化された年齢（40 歳未満／ 40 歳以上）」「二値化された学歴（専門学校以下／短大・高専以上）」「二値化された世帯所得（1,000 万未満／ 1,000 万以上）」「専門的・技術的職業従事者／それ以外」、連続変数としては「年齢」「威信スコア」「クリエイティブ・スコア」である。

予測対象となるそれぞれの変数に対して、例えばカテゴリ変数ならばロジスティック回帰モデルやランダムフォレストなど複数のモデルについて、単純にソフトウェア(Python の scikit-learn)のデフォルト設定によって学習を行い、予測性能を暫定的に比較した。その結果、L2 正則化ロジスティック回帰モデルがほとんどのモデルで有望であったため、その後このモデルについてグリッドサーチクロスバリデーションを行い、最適なハイパーパラメータを探索した後で、あらためてその性能を検証した。結果として、ある程度の精度で予測できたのは、「性別」「40 歳未満／ 40 歳以上」のみであった。カテゴリ変数として、「二値化された学歴（専門学校以下／短大・高専以上）」「二値化された世帯所得（1,000 万未満／ 1,000 万以上）」「専門的・技術的職業従事者／それ以外」、連続変数として「年齢」「威信スコア」「クリエイティブスコア」は、ベンチマークをある程度上回る予測性能をもつ予測モデルを学習できなかった。そこで、以下では、「性別」「40 歳未満／ 40 歳以上」を予測対象とした分析結果について詳細に報告する。

【性別】

まず、サンプルにおける分布を確認すると、男性が 56.3%、女性が 43.7% であった。これに対して、L2 正則化ロジスティック回帰のテストデータにおける正解率は、86.2% であった。ロジスティック回帰モデルの場合には特徴量の貢献度は単純に係数の絶対値の大きさをみることで分かる（表 3 参照）。

第一に、日本語における第一人称の使用の仕方は男性性、女性性を強くシグナルする。また、狭義の文化的実践について述べると、「ラーメン」「ビール」「pepsi」といった食文化は男性性を強くシグナルする。「食う」というボキャブラリも男性に特有である。その他、「車」や「ヨドバシ」のような電化製品も男性性と結びついていることがみてとれる。他方、女性の場合には、「肌」のような美容を思わせる語が上位にきている。第三に、Twitter の一つの特徴として、アイドルや芸能人の応援、いわゆる「推し活」に関わる語が上位にきている。男性の場合には、「チェキ」という実践が上位にきているのは興味を引く。推し活」の文脈では「チェキ会」とはアイドルなどの自分の「推し」と2 ショット写真などを撮影できる機会のことを指す。男性女性の推し問わず、存在する実践であるが、シグナルとしては強く男性性を示唆するようである。これに対して、女性は「くん」という男性向けの呼び名が上位にきているが、これは「推し活」の文脈での呼びかけにも使われる。その他、「推し活」には限定されないものの、「嬉しい」「大好き」のような感情表現も女性については上位にきており、ジェンダーによる感情表出規範の相違とそこから発せられるシグナルが認められる。またこれもやはり「推し活」に限定されないものの、絵文字と思われる「ハート」「大泣き」(neologd は絵文字を言葉に変換する)の使用も女性性をシグナルするようだ。

表 3 性別予測の特徴語上位 20

	男性	女性
1	俺	くん
2	僕	ハート
3	ラーメン	嬉しい
4	車	わたし
5	投資	素敵
6	いく	先生
7	ビール	旦那
8	jal	あたし

9	無い	合掌
10	乃木坂 46	星野源
11	pepsi	肌
12	チャンス	セット
13	勝つ	大好き
14	食う	gt
15	ヨドバシ	商品
16	っす	病院
17	チェキ	ひよこ
18	期待	札幌
19	やはり	大泣き
20	アニメ	涙

【年齢】

年齢については連続値で尋ねているが、これを線形回帰モデルで予測しようとしても全くうまくいかなかった。そこで、より容易な予測問題として、年齢が 40 歳未満か 40 歳以上かという二値カテゴリの予測について検討した。

まず分布については、40 歳未満が 59.4%、40 歳以上が 40.6% となっている。したがって、ベンチマークの正解率は 59.4% である。ここでもやはり L2 正則化ロジスティックモデルが有望であったため、性別の場合と同様に、グリッドサーチクロスバリデーションを行い、ハイパーパラメータ C が 10 のときに、予測性能が最も高く、テストデータでの正解率は 79.3% となった。

特徴語の貢献度を見ると（表 4 参照）、まず言葉遣いに関しては、40 歳未満では、「めちゃくちゃ」「まじ」「やばい」といった言葉が上位にあり、これらは、若者言葉といってよいかもしれない。40 歳以上で語彙というより記号の利用で目につくのは、「艸」「σ」「t_t」「゜д゜」といった記号・漢字である（20 位以内には「艸」だけであるが、50 位まで広げるとその他の記号・漢字もランクインする）。これらは「σ（^_^）」「(T_T)」「(＊´ 艸 `)」「(゜ Д゜)」などの顔文字の一部と思われる。最後に、狭義の文化内容について見ると、40 歳

未満では、「fgo」（「Fate Grand Order」）や「pokemon」などのソーシャルゲームのタイトルがランクインしている。

表 4 年齢予測の特徴語上位 20

	男性	女性
1	僕	応援
2	めちゃくちゃ	不機嫌
3	まじ	娘
4	やばい	ちょっと
5	俺	暑い
6	可愛い	美味しい
7	分かる	下さる
8	fgo	見える
9	キラキラ	息子
10	いく	汗
11	きた	新潟
12	記録	投稿
13	pokemon	艸
14	やつ	怒り
15	くらい	research
16	保育園	今朝
17	チーズ	日
18	終わる	楽天スーパーポイント
19	行く	ヽ
20	ほしい	子

【「二値化された学歴（専門学校以下／短大・高専以上）」「二値化された世帯所得（1,000 万未満／ 1,000 万以上）」「専門的・技術的職業従事者／それ以外」「年齢（連続変数）」「威信スコア」「クリエイティブスコア」】

これらについては、ベンチマークを一定程度、上回る予測性能をもつ予測モデルを学習できなかった。そのため、Twitter での投稿から上記の人口学的・社会経済的属性を読み取ることは困難であると示唆される。

4.3 ツイートから見たブランドへの関心

ツイートからユーザーの個人属性とりわけ社会経済的属性を予測するのは難しいことが示されたが、逆のかたちとして、ツイートに現れるブランドに関する言及をユーザーの人口学的・社会経済的属性によって説明する分析を行った。対象とするブランドは、丸の内ブランドフォーラムのブランド生態系調査（2021 年 10 ～ 11 月実施）で好意または非好意が一定以上あった 131 のブランドである。一度でも言及があったかどうかをロジスティック回帰分析で説明させたところ、比較的メジャーなブランドについて以下のような結果が得られた：

- 性別はほとんどのブランド言及に何らかの影響がある
- 年齢は、特に 20 代が言及を減らすことが多い
- 職業指標のうち「階層」が最低位であるとき言及が増えるブランドが少なくない

社会経済学的属性のなかでは「階層」の低さがブランドへの言及に何らかの影響を与える可能性があること、各スコアの効果は必ずしも線形ではないことが示唆された。

5. おわりに

本研究では、Twitter データとサーベイデータを統合することで、Twitter 上での文化的実践（ツイート）とユーザーの社会的差異（人口学的・社会経済的属性）を結びつけた。社会経済的属性のなかでも職業については、500 近い職業コードを設定して、職業威信、社会的必要性、クリエイティブネスなど多次元的なスコアを独自に算出した。ツイートがユーザーの社会的差異をシグナルしているかという問題設定に応えるべく、疑似予測法の枠組みにより、ツイートからユーザーの人口学的・社会経済的属性を予測できるかを検討した。それなりの精度で予測できたのは、「性別」と「40 歳未満／ 40 歳以上」だけであり、

多くの社会経済的属性については予測できなかった。

そこから示唆されるのは、１つは、ソーシャルメディアでの文化的実践は、Bourdieu 的な枠組みとは異なり、大部分は社会的差異から自由な、少なくともそのような差異のシグナルを抑制するような形で遂行されている可能性である。これは、日本社会には趣味の世界に社会的差異を持ち込むべきではないという規範が歴史的に存在する、といった議論と関連するかもしれない。こうした傾向がソーシャルメディアにおいてより現れやすいかどうかも探求する価値がある。

本研究の結果は、ソーシャルメディア・データをマーケティング・リサーチに活用する、いわゆるソーシャルリスニングにとっても重要な含意を持っている。ツイートから社会経済的属性を含む消費者のプロファイルをかなりの精度で予測するのは難しいという結果は、従来型のサーベイ調査を今後も併用していく必要があることを示唆している。

とはいえ、本研究にはデータ収集上も分析手法上もいくつかの限界や問題がある。それらを解決すべく今後さらに研究を深化させていきたい。

［謝辞］

ブランド生態系調査のデータをご提供いただいた丸の内ブランドフォーラム代表の片平秀貴氏、第１回計算社会科学会大会でコメントいただいた方々に感謝申し上げます。