203-11in

マイクロブログにおける潜在的価値観の推定

Statistical inference of latent values extracted from microblogging

谷田 泰郎[*1] 河本裕輔[*1] 馬場彩子[*1]
Yasuo Tanida Yusuke Kawamoto Ayako Baba

[*1] シナジーマーケティング株式会社
Synergy Marketing, Inc.

The latest developments of digital technologies, such as the explosion of social networking, are inducing rapid and significant changes on consumers' behavior, empowering their choices while lessening the efficiency of classical marketing. However, the wealth of data thus generated offers new opportunities for marketers to analyze consumer psychology. Here we aim at building a statistical model capturing a wide picture of consumer's values and behavior. Specifically, we experiment an inference method predicting a consumer's values from the comments published on its microblog, and discuss issues related to the model's accuracy.

1. はじめに

テクノロジーの進化によるデジタル情報空間の拡大は，消費行動に大きな変化をもたらしている．今日の消費者は，能動的に活動し，お互いに送受信を繰り返すことで社会的ネットワークを構築し，ブランド醸成にまで参画し先導する，まるでアーティストのようである．フィリップ・.コトラーのマーケティング 3.0[1]によれば，協働，文化と共に重要なのが，創造的社会における精神性である．創造性が精神性を刺激し，精神的要求はヒトを突き動かす最大の動機であり，より深いところにある創造性を解き放つ．消費者は，精神を感動させる経験やビジネスモデルを求めており，心理的精神的便益こそが最も基本的なニーズであり，マーケターが実現できる究極の差別化である，と言う考えが示されている．

我々の考え方もこれに近いが，日本においては，ターゲティングやポジショニングと言った従来型のマーケティング手法が無くなるわけではなく，社会に所属するすべての消費者が協働で創るブランドが犇めく将来のマーケットの中でも，その手法は残存していく．但し，活性化する消費者を説明するためには，デモグラフィック属性に加えて，心理的，精神的な属性が必須になる．

このような背景の中，我々は，62 個の価値観成分により説明される社会的類型（以後，本稿では Societas と呼ぶことにする）と情緒ベネフィットを中心に据えて，多くの企業が持つ購買履歴，メール情報，商品やサービスのレビュー，WEB の回遊履歴，ブログやツィート情報，消費者調査情報など，様々なデータとリンクすることによって得られる関係性を，消費者の行動を説明するモデルとして量産して行くための研究をしている[2][3][4][5]．

Societas のような社会的類型を規定した理由の一つは，既に述べたように，マーケットを能動的に徘徊し，社会的ネットワークを構築し，時にはブランディングを行い，マーケット自体でさえ創造してしまいそうな勢いのある消費者を精緻に説明するため，すなわち，マーケターがターゲティングやポジショニングを行う際の消費者像をスピリチュアルに説明することで，マーケティング施策にインスピレーションを与えるためである．

また，Societas，価値観，情緒ベネフィットをマーケットに参加する各種の企業や社会的活動体も含めた法人，個人の枠組みを超えた物差しとして利用することで，社会構成員に集合知を還元することが可能になる．さらには，心理的精神的便益に裏打ちされた行動モデルを量産し，これらの枠組みを精緻化して行くことで，Societas そのものをマーケットの将来予測を行う際のエージェントの複製元として用いることも理論上可能となる．

これらの一連の研究の中で，発言者の言語能力や表現方法も含めた発言そのものを価値観だと考え，マイクロブログ上の発言から心理的属性である価値観を推定する実験を行ってきた[2]．先行研究では，価値観と発言者の言語的成分をモデル化し，Twitter のようなテキスト情報を証拠にした場合の Societas 推定に関する有効性を確認している．本稿では，先行研究のモデルにデモグラフィック属性を加え，言語的成分，デモグラフィック属性，価値観成分，Societas の関係を分析した結果得られた価値観や Societas の推定に影響を与える新たな知見について報告する．

2. 実験に利用した価値観の定義

2.1 価値観の定義

ヒトの持つ潜在的な情報の伝達単位は，ヒトからヒトに複製されながら変化を続けている．このような情報の伝達単位はコミュニケーションの伝達単位であり，コミュニケーションを可能にする複雑な脳を持つヒトのような生物に限定され，動物学的な遺伝子と区別されるコミュニケーションにおける自己複製子は，meme（ミーム）と呼ばれる概念に近いものである．然しながら，我々の目的は，脳の中に定住しているミームを確認することではなく，コミュニケーションの伝達単位を脳の外側で観察できる表現型効果，外界での帰結としてとらえ，マーケティングコミュニケーションのツールとしての精度を向上させることである．本研究が目指す範囲は，消費行動に限定したものであり，その視点から価値観を情報の伝達単位と考え，定性調査と定量調査を実施した[2][3][7][8]．

本稿では，調査の内容や価値観の定義の方法についての詳細は述べない．本稿で利用した価値観に関するデータは，定性調査（インタビュー）によって消費行動における価値観の仮説を立て，2 度の定量調査（WEB アンケート）を行うことで得たものである[2][3][9]．

調査分析の結果，表1に示す 62 個の価値観成分が定義され，本実験では，そのうち Societas への情報量が大きかった 22 個の価値観成分を利用している[2]．22 個に絞った理由は，62 個の成分を作成するためには，303 項目の2水準の選択肢に回答してもらう必要があり，回答者への負担を考慮したことと，後述するマイクロブログと価値観のモデルを作成する際に条件付き確率

連絡先：谷田泰郎，シナジーマーケティング株式会社システム開発部研究開発グループ， 電話番号： 06-4797-2900，メールアドレス： tanida.yasuo@synergy101.jp

表(Conditional Probability Table)が疎になることを避けたかったからである.

価値観フレーム	成分数	成分の内容（ネーミング）
基本的な性格	11	好奇心旺盛　デリケート　マイペース　協調型　勤勉　上昇志向　短気　正義感　ルーズ・不精　無気力　文系的
ポジティブ価値観	8	自己愛　自己成長　アウトドア　スポーツ　恋愛　趣味　ギャンブル　ひとり時間
ネガティブ価値観	3	否定・批判　非常識　期待はずれ
家族関係	7	結婚願望　不仲　責任感（主婦軸）　責任感（扶養軸）　良好（別居家族）　不十分　良好（同居家族）
友人関係	8	ストレス　親友中心　ネットワーク重視　社交的　大人数派　消極的（独身）　仕事人脈中心　ノンストレス
仕事に対する価値観	6	満足　ストレス　プライベート重視　キャリアアップ転職願望　堅実　社会的意義
時間に対する価値観	11	ゆとり　余裕がない　充実　仲間優先　家族優先　趣味優先　インドア派　アウトドア派　家事分担　退屈　自己投資
お金に対する価値観	8	ギリギリ　ゆとり　貯蓄志向　家族優先　慎重派　自己投資　堅実生活　常識的

表１：価値観成分

2.2 Societas(価値観の社会的類型)

被験者ごとの価値観の成分得点の分布を基にクラスタリングを行うことで Societas を規定した[2][3]. 表2に規定した価値観の社会的類型である 12 個の Societas を示す.

#1-1	受け身な 隠者タイプ	常に平常心を保ち，消極的で物静か．周囲に流されなることはあまりない.
#1-2	受け身な 清閑タイプ	物事にあまりこだわりがなく達観しているところがある．その反面，夢中になれるものがあまりないので，趣味を見つけて楽しみたいと思っている．片付けが苦手な面も.
#2-1	家族大好き 悠々タイプ	こだわりや夢中になれるものはあまりなく，少しルーズなところがある．スポーツやアウトドア，趣味を楽しむ面も.
#2-2	家庭的な 真面目タイプ	勤勉で正義感が強いタイプ．アウトドアやスポーツ，趣味を楽しむ.
#3-1	こだわり インドア派タイプ	デリケートだが，正義感がある．恋愛への憧れがあるが，消極的．一人で趣味に没頭する事が楽しい.
#3-2	アウトロータイプ	デリケートで感情の起伏が激しい面も．宝くじなどに当たると嬉しい．整理整頓が苦手.
#4-1	自分中心的な アクティブタイプ	自分磨きや自己成長に余念がない上昇志向タイプ．恋愛にも積極的.
#4-2	好奇心旺盛な バランス人間タイプ	成長意欲が強く，そのために周囲との協調性も大切にする好奇心旺盛な行動派.
#5-1	家族想いの 多忙ワーカータイプ	協調性はあるが自分の価値や時間を大切にするマイペースなタイプ．非常識な事や批判されることが嫌い.
#5-2	社交的な 堅実ホームメーカータイプ	協調性があり，人に喜んでもらう事で幸せを感じる．また，好奇心も旺盛で，自分磨きも怠らない.
#6-1	繊細な 個人主義タイプ	マイペースで，一人の時間を好む．また繊細な心を持ち合わせている．それゆえ，批判などに傷つきやすく，少し感情の起伏が激しい面がある.
#6-2	好奇心旺盛な 人生謳歌タイプ	好奇心旺盛で成長意欲が強い．協調性もあり，恋愛を楽しみたいタイプ.

表２：Societas（価値観成分の社会的類型）

2.3 価値観モデル

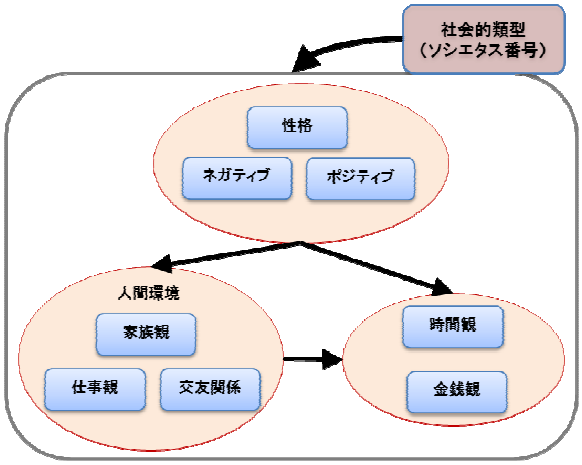


図１：価値観モデル

先行研究にて，被験者ごとに価値観成分(表1)と Societas 番号(表2)を与え，それを教師データとして学習することでベイジアンネットワークによる確率モデル(学習に用いたデータは，11,410 件．価値観成分はすべて「あり／なし」の 2 値に離散化．Societas 番号は 12 水準．構造探索の制約条件は，図1の矢印の先が親になることを許容．円で囲まれた部分のカテゴリ内の変数同士はフリー探索）を構築した[2][3]．次項で述べる Twitter-価値観モデルは，ここで構築した価値観モデルに Twitter 発言者の価値観を証拠として与えることで得られた Societas 番号変数の事後確率を教師データにして構築した.

3. Twitter-価値観モデルを用いた実験

本研究では，発言の取得のし易さ，発言の即時性や匿名性により，消費行動におけるヒトの本音や性質が現れやすいことなどの理由から，分析するマイクロブログの対象として Twitter を採用している．先行研究では，価値観と発言内容を紐づけるための定量調査により取得した基礎データを利用して Societas を言語的成分と価値観で説明するモデルを作成して，言語成分による Societas 推定の可能性を確認した[2]．本稿では，これにデモグラフィック属性を加えてモデルを拡張し，言語的成分及びデモグラフィック属性が，各々単体で，あるいは共起して，価値観推定及び Societas 推定にどのような影響を与えるのかを確認した.

3.1 Twitter-価値観モデルの拡張

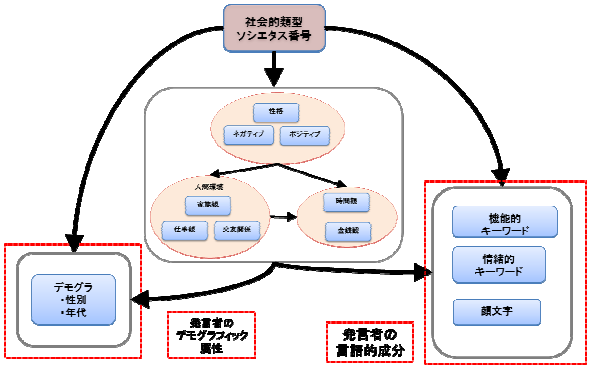


図２：拡張した価値観モデル

先行研究[2]のTwitter-価値観モデルにデモグラフィック属性（性別・年代）を加えて拡張した（図2）．確率モデルの作成方法は，先行研究に準じている．まず，図1に示した価値観モデル[2][3]を用いて，Twitter発言者の価値観成分を証拠として与えた時のSocietas番号の事後確率を求めた．そして，Societas番号（12水準）の事後確率の分布に応じてウェイトバックをかけ（例えば，Societas番号1-2である確率が，17%であれば，1の位を四捨五入して20%とし，そのデータを2件複製，100%であれば10件複製）教師データとした．図2は構造探索の条件である（矢印の先が親になることを許容．円で囲まれた部分のカテゴリ内の変数同士はフリー探索）．ベイジアンネットワークにおける推論アルゴリズムは，精度を優先してMSSMを採用した．また，価値観成分は前述のようにSocietasへの情報量が大きかった22個の価値観成分，言語的成分は，先行研究で利用したものをそのまま用いた（各価値観成分に対して感度が高い語彙から主成分により抽出した第1成分，第2成分の2変数．合計44変数を「あり／なし」の2水準に離散化）．

3.2 評価実験概要

図2に示した確率モデルを用いて，価値観成分変数のみを証拠として与えた場合の Societas 番号等の事後確率を求め，それを基準にした（上限値と見なした）．評価実験では，言語的成分変数のみを証拠にした場合，デモグラフィック属性変数のみを証拠にした場合，言語的成分変数とデモグラフィック属性変数の両方を証拠にした場合を比較した．

精度評価には，以下に示す類似度を用いた．

類似度　＝　Cos（base, eval）

ここで，Cos は，コサイン類似度を計算する関数，base は価値観変数を証拠として与えた場合の評価する目的変数の事後確率の分布，eval は，評価する説明変数を証拠として与えた場合の評価する目的変数の事後確率の分布である．

評価指標を類似度計算ベースにした理由は，Societas 番号の推論結果のように，確率分布で表現されているものを評価する場合，1 位正解率や 3 位正解率のような評価にするより正確に評価できると考えたからである．

具体的には，Societas 番号に対する精度評価をする場合，価値観変数のみを証拠として与えた場合の Societas 番号変数の事後確率の分布と評価対象変数を証拠として与えた場合の Societas 番号の事後確率の分布の類似度を評価指標としている．同様に，価値観に対する精度評価をする場合も，価値観変数のみを証拠として与えた場合の価値観変数の事後確率の分布を基準としている．これらの場合，価値観を証拠として与えた時の類似度は 1 と言うことになる．

尚，評価実験に用いた Twitter 発言者は 562 人である[2]．

3.3 評価結果と考察

Societas 番号の推定に関する評価結果を表3に，価値観に対する評価結果を表4に示す．繰り返しになるが，各表で示されている評価値は，前述の事後確率（推論結果）の分布の類似度である．

Societas 番号の推定に関して，言語的成分変数のみを証拠に与えた場合の類似度は，平均で 0.65 であった．これは，先行研究の Societas 番号の推定に関する評価（1 位正解率を使っている）において，価値観変数を証拠として与えた場合（上限値）に対して，6,7 割の精度が担保できている[2]，と言う知見と整合しており，それを再確認することになった．

	Societas番号	言語的成分	デモグラフィック属性	言語的成分＋デモグラフィック属性
Societas番号	#1-1	0.61	0.29	0.62
	#1-2	0.88	0.10	0.91
	#2-1	0.61	0.36	0.58
	#2-2	0.57	0.48	0.58
	#3-1	0.60	0.50	0.60
	#3-2	0.63	0.44	0.63
	#4-1	0.57	0.38	0.56
	#4-2	0.64	0.38	0.64
	#5-1	0.67	0.35	0.66
	#5-2	0.68	0.36	0.68
	#6-1	0.67	0.50	0.67
	#6-2	0.61	0.43	0.63
コサイン類似度の平均		**0.65**	**0.38**	**0.65**

表３ Societas 番号に対する評価

また，デモグラフィック属性変数のみを証拠として与えた場合の Societas 番号の推定結果は，言語的成分変数のみを証拠に与えた場合に比べて非常に悪い，と言う知見を得た． Societas 番号ごとの評価値を比べてみると，特に，#1-2 に対する推定精度が極端に悪い．逆に言語的成分変数のみを証拠として与えた場合には，#1-2 に対する推定精度は高かった．#1-2 の類型は，受け身ではあるが，物事に拘らず達観しているタイプである．このような類型について，デモグラフィック属性に特徴がなく，言語的成分に特徴が表れている，と言うところが興味深い．その他，#1-1，#4-2，#5-1，#5-2 においても顕著な差が見られた．このうち，#5-1，#5-2 は女性率が高く，Twitter の発言者は男性率が高いため，証拠不足になっているものと思われる．

	価値観	言語的成分	デモグラフィック属性	言語的成分＋デモグラフィック属性
性格	デリケート	0.79	0.69	0.79
	マイペース	0.84	0.75	0.85
	協調型	0.80	0.69	0.80
	好奇心旺盛	0.80	0.70	0.82
家族	結婚願望	0.76	0.71	0.77
	責任感（主婦軸）	0.72	0.66	0.73
	不仲	0.72	0.65	0.73
交友関係	ストレス	0.79	0.71	0.79
	親友	0.81	0.76	0.81
仕事観	ストレス	0.74	0.66	0.74
	やりがい	0.70	0.62	0.71
金銭観	ギリギリ	0.84	0.81	0.84
	ゆとり	0.63	0.51	0.63
	貯蓄志向	0.55	0.42	0.55
ネガティブ	期待はずれ	0.76	0.69	0.75
	否定	0.72	0.65	0.72
	非常識	0.81	0.78	0.81
ポジティブ	自己愛	0.78	0.71	0.79
	自己成長	0.80	0.75	0.81
時間	ゆとり	0.74	0.66	0.74
	家族優先	0.64	0.53	0.66
	余裕がない	0.74	0.65	0.75
コサイン類似度の平均		**0.75**	**0.67**	**0.75**

表４ 価値観に対する評価

価値観に対する評価においても，Societas 番号に対する評価と同様に，デモグラフィック属性変数のみを証拠として与えた時

の価値観の推定結果は，言語的成分変数のみを証拠にした場合に比べて非常に悪い．然しながら，価値観に対する評価では，言語的成分変数を証拠として与えた場合とデモグラフィック属性を証拠に与えた場合での価値観変数の種類による大きな差は見られず，どちらを証拠として与えた場合でも推論しやすい価値観，推論しにくい価値観があると言うことが分かった．性格やポジネガ，金銭的ギリギリや交友関係は Twitter の書き込みからもデモグラフィック属性からも推論しやすい傾向があるようだ．

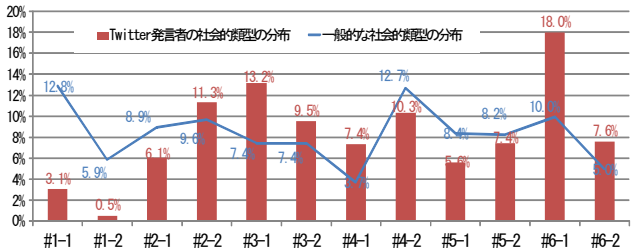
	デモグラフィック属性	価値観成分	言語的成分
性別	性別	0.88	0.78
年代	20才未満	0.53	0.32
	20代	0.73	0.57
	30代	0.71	0.56
	40代	0.71	0.58
	50代	0.61	0.38
	60代	0.37	0.14
	70代	0.50	0.03

表５　デモグラフィック属性に対する評価

また，Societas 番号，価値観に対する評価において共通するのは，言語的成分変数とデモグラフィック属性変数の両方を証拠として与えても，推定精度の向上は見られなかったと言うことだ．その理由を探るために，価値観成分変数を証拠として与えた場合と言語的成分変数を証拠として与えた場合の性別，年代に対する推論結果の評価を行った（表5）．表中の評価値は Twitter のアンケート調査の際に被験者の回答のデモグラフィック属性（性別・年代）を正解とした時の類似度（前出のコサイン類似度を用いた計算方法に準じる）である．デモグラフィック属性変数を証拠にした場合の価値観変数や Societas 番号に対する推定精度はあまり良くなかったが，価値観変数を証拠として与えた場合のデモグラフィック属性に対する推定精度は，0.88 と想定以上に良かった．また，言語的成分変数を証拠として与えた場合のデモグラフィック属性に対する推定精度も，0.78 であり，言語的成分が価値観を推定するためのチューニングをしている（価値観推定に関して情報量が大きい語彙から作成した）ことを考慮すれば，かなり良い数値であると言える．

年代では，20 代～40 代に対する推定精度が比較的良い結果であった．10 代に関しては，基礎調査で取得していない年代であり，また，50 代以上は，Twitter 利用者に少ない年代であるため，データの特性上推定が難しかったのはないかと思われる．

これらの結果を総合的に勘案すると，デモグラフィック属性を加えても価値観や Societas 番号に対する推定精度が向上しない理由は，言語的成分変数のみを証拠とする操作を行うことで，価値観だけでなく，デモグラフィック属性に対する推定精度がある程度担保できているから，と言うことになる．



参考グラフ：Twitter 発言者の Societas 分布

最後に Twitter 発言者の Societas 分布（参考グラフ）を示しておく．青の折れ線が一般（1.1 万人分）の Societas 番号の分布，赤の棒グラフが Twitter 発言者の Societas 番号の分布である．

4.　おわりに

社会的類型や価値観の推定に関して，デモグラフィック属性と言語成分との相乗効果は見られなかったが，逆にデモグラフィック属性が判明しても，価値観や社会的類型と言った心理的精神的な要素を含む類型を推定するのは難しい，と言う知見を得た．さらに，Twitter の発言のようなマイクロブログでの発言から抽出した言語的な成分によって，デモグラフィック属性だけではなくライフスタイルや性格と言った価値観や心理的な属性，社会的な類型を推定できると言うことを確認した．

現時点では，価値観成分や言語的成分を抽出する時のクラスタリングの問題や確率モデルの構造の問題，抽出された成分の妥当性など基礎的な方法論に関する課題も多いが，本稿では利用していない行動履歴，サービスに対するベネフィット，あるいは，フォロワー数やフォロー数，1ツィートあたりの語彙数や推定使用語彙数，プロフィールから取得できる情報などの社会的なネットワーク構造，その他利用可能な履歴情報をリンクしていく必要がある．我々の大きな目的は，社会の構成員である法人や個人の協働作業によって還元可能な社会知ネットワークを形成していくことであり，そのためにも，将来に向かって拡大しつつある社会知ネットワークの利用手法を確立し，様々な側面での人的，社会的なマーケティング活動を支える行動予測モデルを量産して行くことが重要だと考えている．

参考文献

[1]　P.コトラー，H.カルタジャヤ，I.セティアワン(2010)．恩蔵直人監訳．藤井清美訳．コトラーのマーケティング 3.0．朝日新聞出版

[2] 谷田泰郎，馬場彩子，河本裕輔，藤井絵美子(2013)．価値観モデルを利用したマイクロブログ発言者の社会的類型の推定．言語処理学会第 19 回年次大会(NLP2013)

[3] 馬場彩子，谷田泰郎，Bertin Mathieu(2013)．社会知としての消費者価値観構造モデルと類型「Societas」の構築．人工知能学会全国大会（第 27 回）JSAI2013

[4] 木虎直樹，久保 征人(2013)．Web アクセス履歴に基づくユーザの価値観の類推 ．人工知能学会全国大会（第 27 回）JSAI2013

[5] 馬場彩子，木虎直樹，谷田泰郎，後迫彰，井上哲浩，加藤卓(2012)．社会知を還元するクラウド型データベース「iNSIGHTBOX」の構築．平成24年度情報処理学会関西支部大会

[6] 荒牧英治，増川佐知子，森田瑞樹，保田祥(2012)．日本人のオンライン・コミュニケーション上での平均使用語彙数は 8,000 語である．情報処理学会研究報告自然言語処理(NL),2012-NL-208(9)

[7] R・ドーキンス(1987)，日高敏隆他訳．利己的な遺伝子．紀伊国屋書店

[8] R・ドーキンス(1976)，日高敏隆他訳．延長された表現型．紀伊国屋書店

[9] 池尾恭一，青木幸弘，南知恵子，井上哲浩(2010)．マーケティング．有斐閣

[10] 奥村学(2012)．マイクロブログマイニングの現在．電子情報通信学会第3回集合知シンポジウム

[11] 小林哲郎(2012)．ソーシャルメディアと分断する社会的リアリティ．人工知能学会誌　Vor27 No.1

[12] 本村陽一（2006）．ベイジアンネットワーク技術．東京電機大学出版局

[13]　池田和史，服部元，松本一則，小野智弘，東野輝夫(2011)．マーケット分析のための Twitter 投稿者プロフィール推定手法，「マルチメディア，分散，協調とモバイル（DICOMO2011）シンポジウム」