Vol.2014-IFAT-114 No.7
Vol.2014-DD-93 No.7
2014/3/29

ユーザの嗜好と所有物の関係性を用いた属性分析

馬繰 美穂[1,2,a)]　徳久 良子[2,b)]　寺嶌 立太[2,c)]

概要：近年，ユーザの意見を直接収集できるソーシャルメディアを商品の企画・設計・開発などに活用することが検討されている．しかし，ソーシャルメディアは発信者がプロフィール情報を明示しない場合も多いため，十分に活用できていないのが現状である．本稿では，ソーシャルメディア上のテキストからユーザのプロフィール情報の 1 つである趣味嗜好を推定する手法を提案し，その有効性を評価する．本手法の特徴は，所有物を表す表現を手がかりとして，趣味嗜好とその関連語の知識を低コストで構築した点にある．この知識とソーシャルメディア上のテキストを比較することで，ある趣味嗜好に関連する発言が多いユーザはその趣味嗜好を持つと判定できる．本手法を用いて，Twitter に投稿したユーザの趣味嗜好を推定した結果，プロフィールを明示していないユーザに対しても趣味嗜好を推定できた．

1. はじめに

近年，ソーシャルメディアの流行に伴い，商品の企画・設計・開発などにおいてソーシャルメディアの活用が検討されている．従来の新商品の開発では，アンケートやフォーカスグループインタビューなど「統制のとれた閉じた顧客の声」を集める方法がとられてきたが，ソーシャルメディアを活用することで「市場の生の声」を集めることが可能となり，ひいては市場のニーズをより反映した商品開発が実現できると期待されている．例えば，ソーシャルメディア上で「赤い車がほしい」と発言するユーザの中に，「自転車が好きな」ユーザが多いことがわかれば，トランクに自転車が積める赤い車を設計するなど，ソーシャルメディアにみられるユーザの傾向を把握することで市場のニーズに合った商品を企画・設計・開発できると期待される．

しかし，その一方で，ソーシャルメディアには発信者のプロフィール情報（性別・年代・職業・趣味嗜好など）が不足しているため，現状ではソーシャルメディアを商品の企画・設計・開発には十分活用できていないという指摘もある [1]．プロフィール情報の中でも趣味嗜好（e.g., 自転車が好き）はユーザの消費活動と密接に結びつくため，商品を企画・設計・開発する上で非常に重要な要素と考えられる．

[1] 東京工業大学総合理工学研究科
Tokyo Institute of Technology
[2] 豊田中央研究所
Toyota Central R&D Labs., Inc.
[a)] matsunag@lr.pi.titech.ac.jp
[b)] tokuhisa@mosk.tytlabs.co.jp
[c)] ryuta@dii.tytlabs.co.jp

そこで本稿では，ソーシャルメディア上のテキストからユーザの属性の 1 つである趣味嗜好を推定し，その有効性を評価した結果を報告する．本手法の特徴は，所有物を表す表現を手がかりとして，趣味嗜好と関連する語をコーパスから効率的に収集することにより，趣味嗜好とその関連語の知識を低コストで構築した点にある．この知識をソーシャルメディア上のテキストに適用することで，ユーザの属性を推定する．以下，2 節では関連研究を述べ，3 節以降で具体的な手法と実験結果を説明する．

2. 関連研究

ソーシャルメディアの情報を用いて，性別や年代などの属性を推定する研究は数多く存在する [1], [2], [3]．これらを大別すると，a) 属性推定の課題を分類問題として定式化し属性を推定した研究と，b) 特定の表現と属性の関係に着目し属性を推定した研究とに分けられる．

a) の先行研究には，Rao ら [2] や平野ら [3] の研究がある．Rao らは性別，年齢，出身地，政治的な信条について，ユーザ間のネットワークなどを素性として，機械学習により属性を判別した．また，平野らは，「年齢が 10 代」であれば「飲酒はしない」のような，属性間に存在する推論関係を学習することで属性の推定を行った．これらの研究は比較的高精度で属性推定が行えるが，学習データの構築にコストがかかるという問題がある．また，性別，年齢，出身地のようにカテゴリ数が限定される場合は分類問題として定式化しやすいが，今回の対象としている趣味嗜好の推定はカテゴリ数が明確でないため（i.e. 性別は「男性」「女性」とカテゴリが限定されるが，趣味嗜好の場合は「自転

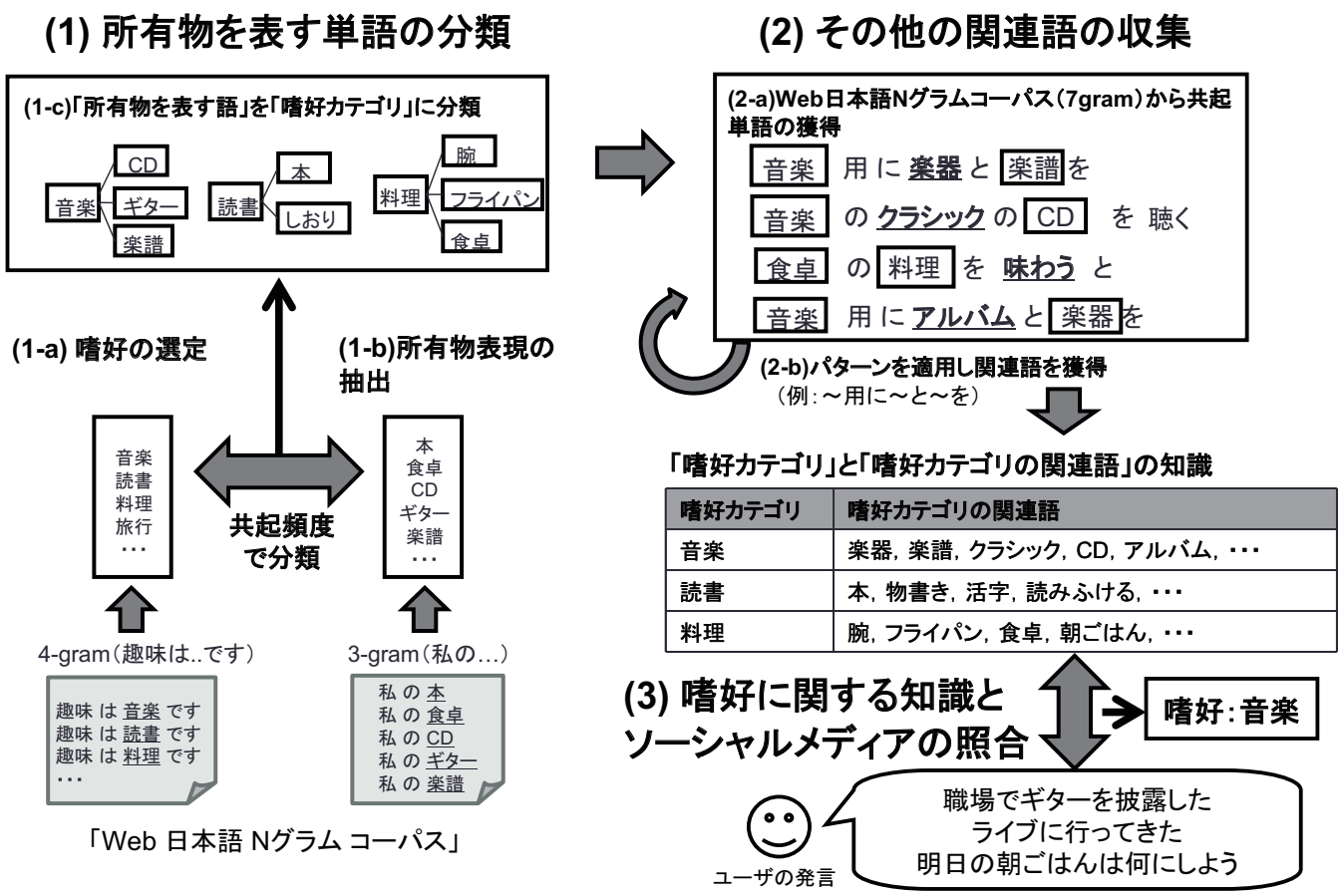

図 1 提案手法の概要

車が好き」「スキーが好き」「プログラミングが好き」「パソコンが好き」など，有限のカテゴリを定義することができない），人手で学習データを構築するのはコストの面から現実的でない．

一方，b) の先行研究として，那須川らは Twitter のデータから一人称所有格の後に現れる名詞を収集することで，その名詞がユーザと特定の関係を持つことを示している [1]．具体的には，「私の車」「自分の車」「うちの車」など一人称所有格の後に「車」が続く形で Twitter に投稿しているユーザは車の所有者である可能性が高いことを調査の結果明らかにし，一人称所有格に続く名詞は所有物としての性質を持つことを示している．また，これを応用することで，車を所有していることなど特定のプロフィールを持つユーザの嗜好や動向を効率的に調査できる可能性を示唆している．彼らの提案は，ソーシャルメディアを活用した商品の企画・設計・開発に非常に有効と考えられる．

しかし，那須川らの研究では属性推定が行われていないため，本稿ではこの知見を用いて属性推定を行う．一人称所有格の後に続く名詞を所有物表現として収集し，嗜好を表す属性と関連付けることで，嗜好推定への手がかりを得ることを目指す．

3. 提案手法

3.1 手法の概要

ある嗜好を持つユーザはその嗜好に関連する語句を多く発言すると考える．例えば「今日は職場でギター弾いてきた」や「ライブに行ってきた」などと発言するユーザは「ギター」「ライブ」などの単語から音楽が嗜好であると推定できる．つまり，「音楽」に対する「ギター，ライブ」のように，嗜好と関連する語の知識を収集できれば，この知識をソーシャルメディアの投稿と照合することで嗜好が推定できると考えられる．

嗜好の関連語を効率的に収集するために，那須川らが示したような，一人称所有格の後に現れる名詞は所有物を表すという関係を用いる．例えば，音楽を嗜好する人間であれば楽器や楽譜を所有しているように，嗜好は所有物と強い関係を持つと考えられる．したがって，もし所有物を表す語がどの嗜好と関連しているかが分かれば，嗜好を推定する手がかりになると期待される．そこで，本稿では「私の」に続く名詞を収集し，嗜好と関連づけていく．

しかし，性別や年齢などとは異なり，嗜好を表す語の集合は自明ではない．そこで，どのような嗜好や所有物が実際に存在するかを調べるために，Google が提供する Web

旅行, 読書, 写真, 料理, 音楽, ゴルフ
ドライブ, パチンコ, サッカー, バイク
ゲーム, コンピュータ, スキー, 生け花

図 2 嗜好表現の一覧

楽譜, 家, 名前, サイト, グローブ
おみやげ, レシピ, モーグル, 競技
オートバイ, かばん, ブログ

図 3 所有物表現の一覧

日本語 N グラムコーパス [4] に，後述する特定のパターンを適用して嗜好や所有物を抽出する．Web から収集された約 250 億文からなる大規模なコーパスを利用することで，複数の単語で構成されるパターンでも十分な頻度の情報を得ることができると考えられる．

嗜好カテゴリとその関連語の知識を構築し，嗜好推定を行う処理を図 1 に示す．以下では詳細について説明する．

3.2 嗜好と所有物表現の選定

まず，推定対象となる嗜好を選定する．選定候補としては「音楽」「自転車」「スキー」など様々な嗜好が考えられるが，嗜好を表す語としてどのような語を選定すべきかは自明ではない．そこで，図 1 の (1-a) に示すように，Web 日本語 N グラムコーパスの 4 グラムコーパスから「趣味は○○です（○○には自立語の名詞が入る）」というパターンにマッチする単語を抽出し，「○○」の部分の語を嗜好を表す語とした．「趣味は○○です」という複数の語からなるスパースな表現を用いるものの，大規模な Web コーパスから抽出することで必要な量は確保できると考える．図 2 に抽出した単語の一部を示す．

また，所有物を表す単語については，図 1 の (1-b) に示すように，Web 日本語 N グラムコーパスの 3 グラムコーパスから「私の○○（○○には自立語の名詞が入る）」というパターンを抽出し，「○○」の部分にマッチする語を所有物を表す語とした．図 3 に抽出した単語の一部を示す．

3.3 所有物と嗜好の関連付け

3.2 節では嗜好と所有物の表現を抽出する方法を述べたが，嗜好と所有物の表現の間の関係については明らかではない．これらの表現を属性推定の手がかりとして利用するためには，所有物表現がどの嗜好と関連するかを決定する必要がある．本節では，図 1 の (1-c) に示すように，嗜好カテゴリと所有物表現の共起頻度によって嗜好と所有物表現を結びつける方法を説明する．

具体的には，Web 日本語 N グラムコーパスの 7 グラムにおいて嗜好カテゴリと所有物表現の共起頻度を求め，所有物ごとに最も頻度が高い嗜好を関連づけた．ただし，嗜好の出現頻度に差があるため，単純に共起頻度を比較することはできない．したがって，共起頻度を嗜好カテゴリの語の出現頻度で正規化した値を用いることで，該当する所有物表現が嗜好と共起する単語に占める割合を関連度の比較基準とした．

嗜好カテゴリ	所有物表現
音楽	楽譜, 打楽器, レコード, ギター
料理	焼鳥, 小皿, レシピ, ダイニング
読書	書斎, 感想, コミック, しおり

表 1 嗜好と所有物表現の対応

また，3.1 節で抽出されたすべての嗜好を表す語について所有物表現を分類するのではなく，今回は試みとして，嗜好を表す語のうち頻度の高かった 12 単語に限定し所有物表現との関連度を調べた．この処理によって，それぞれの所有物表現が最も関連する嗜好に分類される．このようにして構築した嗜好カテゴリと所有物表現の対応の一部を図 1 に示す．

3.4 パターンを用いた関連語の獲得

3.3 節では，嗜好カテゴリと所有物表現の知識を構築する方法として所有物表現は嗜好と関係が深いというヒューリスティクスを利用することで，効率的に嗜好カテゴリの関連語を獲得した．しかし，このままでは嗜好カテゴリの関連語が所有物を表す名詞に限定されているため，本節では動詞を含むその他の関連語を収集することを目指す．

まずはじめに，図 1 の (2-a) で示す通り，Web 日本語 N グラムコーパスの 7 グラムコーパスにおいて嗜好カテゴリと所有物表現が共起する文を抽出し，これらの語と共起する単語を関連語として収集する．この時，コーパスの文から嗜好カテゴリと関連語を除くと「○○用に××と□□を」や「○○の××の□□」のようなパターンが抽出できる，ここで，○○は嗜好カテゴリ，□□は所有物表現や既に獲得されている関連語，××は獲得したい関連語を指す．次に，図 1 の (2-b) のようにこのパターンを再び 7 グラムコーパスに適用することにより，更に関連語を収集する．

例えば，3.3 節の方法で「音楽」という嗜好に対して「CD, ギター，楽譜」という所有物表現が分類された場合，図 1 の (2-a) の 1 行目にあるような「音楽」と「楽譜」が共起する「音楽用に楽器と楽譜を」が最初に抽出され，新たな関連語として「楽器」が獲得される．更に「○○用に××と□□を」というパターンをコーパスに再び適用することで，図 1 の (2-a) の 4 行目「音楽用にアルバムと楽譜を」がマッチし，新たな関連語として「アルバム」を得る．

嗜好カテゴリ	嗜好カテゴリの関連語
音楽	ギター，楽譜，引っさげる，耳慣れる，クラシック，軽音楽，原盤，楽器，音質，シンセサイザー，アーティスト，新譜，エレクトロニクス
パソコン	機器,モジュラ，取りこむ,組み立る，ノートブック，自作，プリンター，ディスプレイ，ウィンドウズ，モデム，電源，接続，インターフェース
料理	焼き豆腐，手が込む，炙る，重詰，めしあがる，エスニック，オイスター，つけあわせる，手ぬき，ヘルシー，引立てる，ふきこぼれる
サッカー	ハイビジョン，天皇杯，スパイク，ワールドカップ，イングランド，エクアドル，ロスタイム，ロングパス，ヘディング，蹴球

表 2 嗜好カテゴリと関連語

以上の方法により，最終的に嗜好 12 カテゴリに対して約 14 万 1 千語の単語を獲得した．表 2 に獲得した単語の例を示す．例えば「音楽」の関連語として「引っさげる」「耳慣れる」などのような動詞も獲得できている．

3.5 ソーシャルメディアへの適用

図 1 の (3) に示す通り，3.4 節までに構築した嗜好カテゴリとその関連語の知識を，ソーシャルメディアの投稿内容と照合することで嗜好を推定する．方法の詳細を図 4 に示す．

図 4 の (1) に示すようにユーザのソーシャルメディアへの投稿の中から内容語を抽出して出現頻度を成分としたベクトルを作成し，嗜好カテゴリとその関連語の知識との類似度を求める．類似度計算の際は，3.4 節までに求めた共起頻度を各関連語の Web 日本語 1 グラムコーパスの頻度で割ることで正規化し，関連語のスコアとする．今回はベクトル間の類似度を求めるため尺度としてコサイン類似度を用いる．

次に，最も類似度の高い嗜好カテゴリを選択するが，単純に各嗜好カテゴリのコサイン類似度を比較することはできない．なぜなら，例えば，料理が趣味でない人でもソーシャルメディア上ではごはんなど料理に関する内容を話題にする人が多いが，音楽については音楽を趣味とする人しか話題にしないといったように，嗜好カテゴリによってソーシャルメディア上での話題になる傾向の強さが異なるためである．そこで， 図 4 の (2) に示すように，嗜好ごとにユーザの投稿とのコサイン類似度をランク付けし，ランク上位のユーザはその嗜好を持つと判定する．

4. ソーシャルメディアを用いた嗜好推定実験

本節では，手法を評価するための実験及びその結果について述べる．3 節で述べた手法を実際に Twitter のデータに適用してユーザの嗜好推定を行い，上位のユーザを人手で評価する．

4.1 実験データと実験方法

実験データとして，TwitterAPI[*1]を用いてクローリングした 2013 年 5 月 6 日から 7 月 6 日までの 2 ヶ月分のツイートデータから，20 ツイート以上が取得できた 1800 名の約 60,000 ツイートを利用した．これらのデータに対して 3 節までに述べた提案手法を適用し，各ユーザに対して嗜好を付与した．各嗜好ごとに，ユーザとのコサイン類似度をランク付けし，上位のユーザはその嗜好を持っていると判定する．今回は，12 種類の嗜好カテゴリについて上位 10 名のユーザを選定し，判定された嗜好と実際の嗜好の一致する度合いを 2 人の評価者が独立に評価した．

評価は次のような 4 段階で行った．

- ◎：人手で判断した嗜好が判定された嗜好と一致した
- ○：人手で判断した嗜好が判定された嗜好と概ね一致したとみなせる
- ×：人手で判断した嗜好が判定された嗜好と一致しなかった
- △：人手では嗜好を判定できなかった

商品の企画・設計・開発に向けてソーシャルメディアからユーザの嗜好を推定するという目的では，再現率も非常に重要となる．しかし，Twitter に存在するユーザの嗜好全体を把握することは困難であるため，今回は判定したデータ中に占める正解の割合を示す適合率によって評価する．

	評価者1	評価者2
◎のみ正解	35.8%(43/120)	25.0%(30/120)
◎＋○を正解	51.7%(62/120)	39.2%(47/120)

表 3 実験結果 (適合率)

4.2 実験結果

表 3 に各評価者ごとに算出された提案手法の適合率を評価個数と共に示す．表 3 の◎は「人手で判断した嗜好が判定された嗜好と一致した」もののみを正解とした場合，表 3 の◎＋○は「人手で判断した嗜好が判定された嗜好と一致した」および「人手で判断した嗜好が判定された嗜好と概ね一致したとみなせる」ものを正解とした場合を表す．

なお，図 3 の評価値には評価者によって差があるが，これは両者の評価基準の差によるものである．嗜好の判定に評価者間で差があった例としては，次のような例となる．

パソコンもう一台ほちぃ

この事例に対して，評価者 1 は「パソコンがもう一台欲しいと言っているからパソコンが趣味だろう」と解釈して○を付与した．しかし，評価者 2 は「パソコンを趣味とするならもう少しパソコンについて詳しく言及するだろう」と解釈し，さらに他の発言からも嗜好が明確に判断できなかったため△を付与した．このように嗜好とみなす基準には個人差があるため，評価結果に差が見られた．

なお，図 3 の結果の精度については必ずしも十分とは言

[*1] https://dev.twitter.com/docs/streaming-api

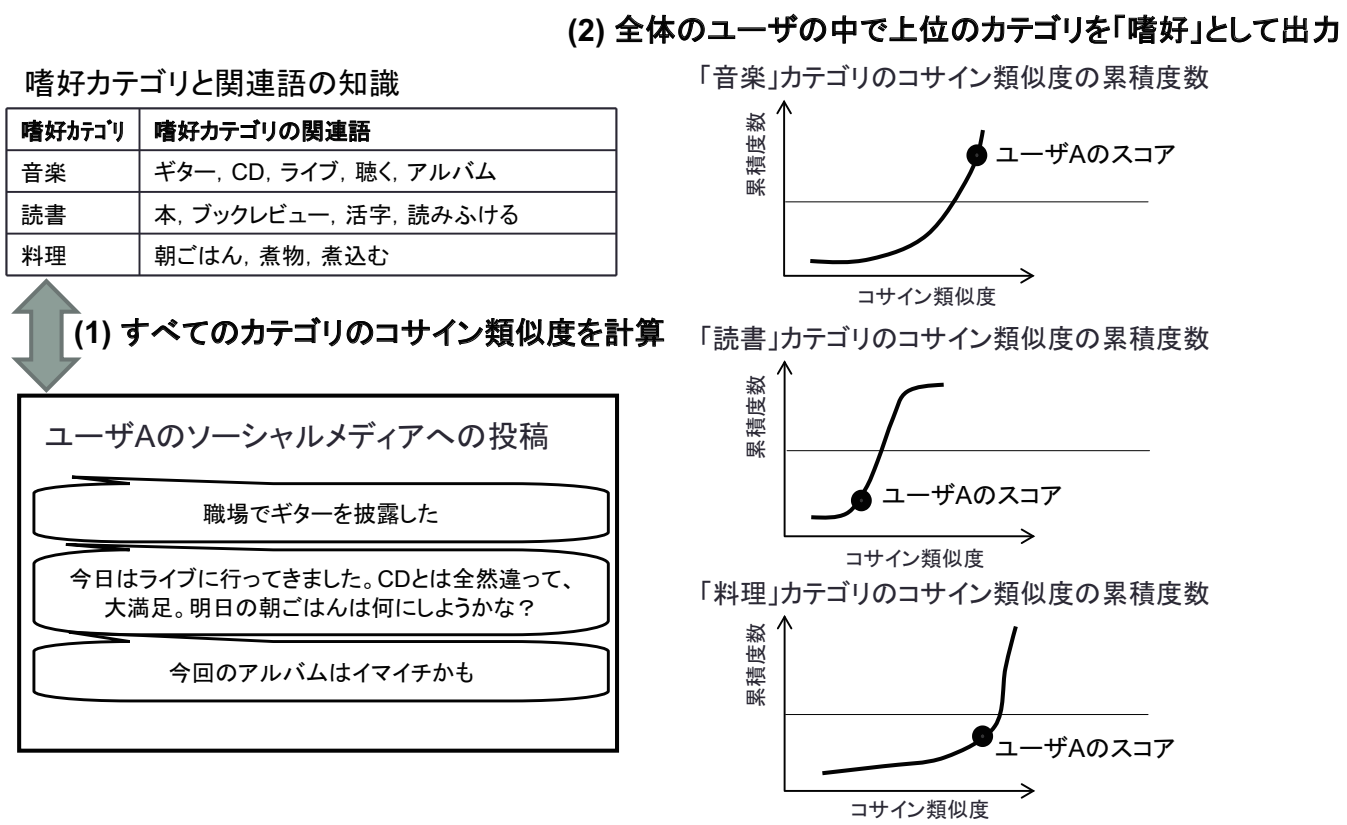

図 4 ソーシャルメディアからの嗜好推定

えないが，今後の改善案に関しては次節で議論する．

5. 議論と今後の改善案

今回，いくつかの例については直接的に嗜好に言及していない場合でも嗜好を推定することができた．以下の例では，「写真が趣味だ」とは明示されなくても，嗜好と関連が強い「一眼レフ」という語に着目することで正しく嗜好が「写真」だと判定できた．

そろそろ一眼レフ、きちんとしたの買わないとなぁ…。

ただし，推定に失敗した例も見られた．考えられるいくつかの原因を具体例と共に挙げる．

まず，今回の実験で正しく推定できなかった事例の中には，否定表現を伴う事例が複数見られた．

あああパソコンきらい（´・ω・‘）

このユーザに対して今回のシステムでは「パソコン」と嗜好を推定した．しかし実際には「パソコンがきらい」と述べられており明らかにパソコンは嗜好ではないと考えられる．今回はソーシャルメディアの投稿と嗜好カテゴリに関連する語の知識の照合を単純な bag-of-words で行ったが，「きらい」などの否定の意をもつ表現は述語とその格要素を正しく解析する必要がある．今後は解析単位を節や文などに拡張し，[5] で考察されたような真偽判断の解析も併せて行うことで，否定的な表現を解析する．

また，投稿中で言及されている行為が発言者による行為かどうかが検証できていないために発生した誤りも存在した．以下の例で「@XXXXX」はツイートがユーザ ID が XXXXX である他者に向けた発言であることを指す．

@XXXXX サッカー見る気満々でワロタ w

このユーザの嗜好は「サッカー」と推定された．しかし，このツイートでは「サッカー」の後に「見る気満々でワロタ」と相手の行動を揶揄するような発言が続いていることから，「サッカーを見る」行為の主体は発言したユーザではなくユーザがメッセージを送った相手であるため，ツイートの発信者がサッカーを嗜好としているかどうかは決定できない．このような誤りも前事例と同様に，知識の照合が単語単位に限定されていたことが原因として挙げられる．今後は，節や文などより広い解析単位において，係り受け関係を考慮することで文中における行為者を正しく解析し，対象ユーザ自身の行為から嗜好を推定したい．

最後に，今回は嗜好のみを解析対象としたが年齢や職業など他の属性も合わせて考慮することで解決できる見込みがある例も見られた．ある 1 人のユーザによるいくつかのツイートを以下に示す．

グルコースとかそこら辺の生物毎回 50 分睡眠してたから全くわからない

数 2 やばい www 平均点 43 点

明日妹の運動会（。・ω・。）久しぶりに中学行く

このユーザに対してシステムは嗜好が「ドライブ」であると推定した．しかし，高校の授業科目を表す「数 2」な

どの語からツイートの発言者は高校生であると考えられるため，通常の高校生が嗜好とする可能性が低いドライブをユーザが嗜好としている可能性は低い．今後は嗜好以外の属性も考慮した上で，推定された属性同士に矛盾がないか検証することで，より精度の高い推定を行いたい．

6. おわりに

今回は嗜好と関連付けられた手がかり語を収集することで，ソーシャルメディアのテキストからユーザの嗜好推定を行い，その有効性を評価・分析した．その結果，嗜好を明示していないユーザでも正しく推定が行えた例も存在したが，着目していたのが単語レベルにとどまったことで解析を誤った例も見られた．今後は節や文など，より大きい構造についても考慮していきたい．

また，ソーシャルメディアを用いたマーケティングを行うためには，嗜好に限らず職業などについても同様の推定が有益である．今後，解析対象となる属性を拡大しユーザのプロフィールに関する詳細な情報を提供できれば，ユーザのニーズにより適合した商品の企画・設計・開発に寄与するものと考えられる．

参考文献

[1] 那須川哲哉，西山莉紗，金山博，吉田一星，大野正樹：一人称所有格を用いたプロフィール推定，言語処理学会第 19 回年次大会発表論文集 (2013).
[2] Rao, D., Yarowsky, D., Shreevats, A. and Gupta, M.: Classifying Latent User Attributes in Twitter, *Proceedings of the 2Nd International Workshop on Search and Mining User-generated Contents*, SMUC '10, Vol. 2, New York, NY, USA, ACM, pp. 37–44 (online), DOI: 10.1145/1871985.1871993 (2010).
[3] 平野徹，牧野俊朗，松尾義博：Markov Logic を用いたテキストからのユーザ属性推定，人工知能学会全国大会論文集 (2013).
[4] 工藤拓，賀沢秀人：Web 日本語 N グラム第 1 版 (2007).
[5] 成田和弥，水野淳太，乾健太郎：日本語事実性解析課題の経験的分析，情報処理学会研究報告. 自然言語処理研究会報告， Vol. 2011, No. 17, pp. 1–8 (2011).