Vol.2017-ICS-188 No.1
2017/7/20

ソーシャルメディアユーザの属性推定と実サービスへの活用

榊 剛史[2,1,a)]

概要：社会学やマーケティングにおいて，人の属性（デモグラフィクス）は分析の手掛かりとして重要な位置づけを担っている．ソーシャルメディアデータをマーケティングや社会調査として活用する際にも，ソーシャルメディアユーザの属性を軸に分析することで，その応用可能性は大きく広がる．本論文では，大規模にソーシャルメディアユーザの属性を推定する際に用いた手法について説明すると共に，実際に得られたユーザ属性を分析に活用した事例について紹介する．

Social Media User Attirbute Estimation and Its Application to Real Services

Sakaki Takeshi[2,1,a)]

1. はじめに

従来，社会学や経済学，ビジネスにおけるマーケティングなどにおいて，調査対象を俯瞰・分析するために，個人の属性を用いて調査対象である不特定多数をグルーピングするアプローチはよく用いられている．例えば，社会調査においては性別や居住地，年代，人種，宗教，家族構成ごとにユーザをセグメント化した上で，各セグメントごとに傾向や分析（例えば，層別抽出法など）が行われる．実際，政府や研究機関による社会調査データは，様々な個人の属性ごとに調査結果を分類・集計した形で公開されている．マーケティングにおいても，消費者の年齢や居住地などを用いて消費者行動分析や売上予測などが行われている．このように，個人の属性が調査データのセグメンテーションに用いられるのは，「同じ属性を持つ集団は，同じような行動特性を持つ」という仮定に基づいている．実際，様々な局面においてこの仮定は有効である．例えば，幸福度調査や選挙予測，世論調査，消費者行動予測など様々な社会現象を分析する上で，個人のデモグラフィクスが分析の軸として用いられてきた．

一方，近年，同じく社会学やマーケティングにおいて，ソーシャルメディア上のデータを分析することで，様々な社会現象―商品・サービスに対する評判や政党の支持率，様々なものごとに対する意見など―を観測しようというアプローチが行われている．近年新たに勃興しつつある分野の一つである計算社会科学においては，このようなアプローチは主要なアプローチの一つである．このアプローチにおいても，ソーシャルメディアユーザの属性を知ることができれば，既存の社会学・マーケティングと同様の分析・集計手法を適用することができる．

ソーシャルメディアユーザの属性について，Facebook を始めとする実名利用が前提とされているメディアにおいては，性別，地域，年代といったユーザ属性がすでに登録されているため，容易に既存の分析・集計手法を適用することができる．実際，Facebook においては豊富なユーザ属性情報を利用した広告ターゲティングが広告サービスとして強い訴求ポイントとなっている．しかし，Twitter を始めとする匿名性が高いソーシャルメディアにおいてはこの限りではない．これらのソーシャルメディアにおいては，何らかの手掛かりからユーザの属性を推定する必要がある．

また，ソーシャルメディアの特徴の一つとして，居住地域といった空間的な近接性や性別・年代といった社会関係

[1] 東京大学
The University of Tokyo, 7-3-1, Hongo, Bunkyo-ku, Tokyo 113–8654, Japan
[2] 株式会社ホットリンク
Hottolink, 7-3-1, Fujimi Duplexbiz Bldg.5F, 1-3-11 Fujimi-cho, Chiyoda- ku,Tokyo 102-0071 JAPAN
[a)] t.sakaki@hottolink.co.jp

性的な近接性には縛られずにコミュニケーションができる点がある．この点は特に匿名性が高いソーシャルメディアで顕著に現れると推定される．実際、Twitter 上では，アニメ好きやゲーム好き，ジャニーズファンなど特定の興味を持ったユーザ集合が活発にコミュニケーション・情報拡散している事例がよく見受けられる．これは言い換えれば，既存のユーザ属性（性別，地域，年代）とは関係なく，「共通した興味を持つために同じような行動特性を持つ集団」が生じていると考えられる．つまり，ソーシャルメディアデータを活用したマーケティングや社会学においては，興味やソーシャルメディア上のコミュニケーションに基づいた新たなユーザ属性を構築する必要があると考えられる．

本論文では，下記 2 つのタスクについて，用いた手法を紹介しつつ，実ビジネスに活用した事例を紹介する．

- Twitter ユーザに対して，その発言・プロフィールから既存の個人属性（性別，年代）を推定する
- Twitter ユーザに対して，ソーシャルメディア上のインタラクション行動に基づいて新たな属性を推定する

本論文では, データ取得の容易性，日本でのユーザ規模の観点から，対象とするソーシャルメディアとして Twitter を用いる．ここで紹介する手法はいずれも技術的には新規性はなく，ありふれた手法である．本論文の貢献は，技術的に新規性がある手法を紹介することではなく，下記の 2 点である．

- 一般的な手法を実データに適用することで，ビジネス上で必要な精度・カバー率が得られるかどうかを検証する
- 一般的な手法で得られた結果を、どのように実ビジネスに活用することができるかを提示する

2. 関連研究

ソーシャルメディアからユーザのプロフィールを推定する研究はこれまでも行われている．以下にその概要を述べる．

ユーザの居住地推定に関して，Twitter のユーザはそれぞれプロフィールを設定しており，居住地についても記載する欄が設けられている．Hecht らはそれらを用いて，州単位，市単位でユーザの居住地を推定する手法を提案している [4]．これによれば，約 4 割程度のユーザは州単位で，26%のユーザは市単位で居住地推定が可能であるとしている．Cheng らは，ツイート集合からその土地固有のキーワードを抽出し，確率モデルを生成することで，51%のユーザの居住地を推定することができたとしている [3]．また Lars らは，距離が近いほど友人になりやすいという仮説に基づいて，Facebook の友人関係を用いてユーザの居住地を推定し，70%近い適合率を得ている [1]．

ソーシャルメディアから性別を推定する研究について，Burger らはツイート中の単語や文字列を手がかりに性別を推定する手法を提案している [2]．彼らによればツイートの量や Twitter ユーザプロフィール内のメタデータ量が推定精度に影響を与えるとしている．性別，年齢，宗教，政治的志向の 4 つの属性を推定する手法として，Rao らはフォロワー数やツイート内容，RT 頻度などを素性として，機械学習を用いる手法を提案している [7]．その他にもソーシャルメディアからユーザの属性を推定する研究は存在する [5], [8]．

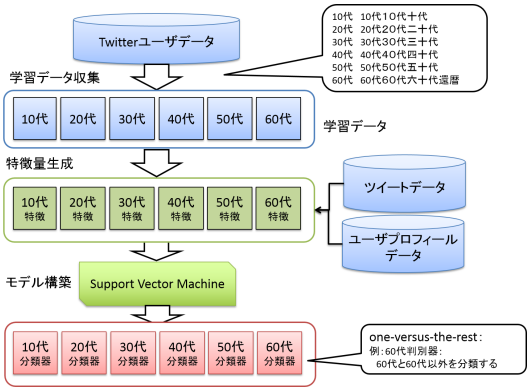

図 1 ユーザ属性推定モデルの構築

本研究では、既存個人属性を推定するタスクに対して，池田らの手法を適用する [8]．また，興味ベースの個人属性を推定するタスクに対して，鳥海・榊らの手法を適用する [9]．

3. 適用手法

3.1 Twitter データを用いた既存個人属性推定

3.1.1 SVM による投稿文・プロフィール文からのユーザ属性推定

本論文では，池田らの手法を用いてソーシャルメディアユーザの個人属性推定を行う [8]．まず，プロフィール文から性別や年代が判定できるユーザを正解データとして収集する．収集したユーザごとに投稿を N_{th} 件以上とプロフィール文を収集し，そこに出現する単語を特徴量として抽出する．そして，収集した特徴量と正解ラベルに Support Vecotr Machine（以下，SVM）用いて，属性推定分類器を生成した．具体的に生成した分類器は，ユーザを男・女に分類する 2 クラス分類器，ユーザを 10 代，20 代，30 代，40 代，50 代，60 代の 6 クラス分類器である．6 クラス分類器では，one-versus-the-rest のアプローチを採用し，各年代ごとに分類器を生成した．分類器の構築方法の手順を図 1 に示す．

3.1.2 学習データの収集

学習用データを収集するために，性別・年代が推定できるプロフィールを抽出するためのルールを人手で整備し，当社内に蓄積された Twitter ユーザデータのプロフィール文に適用し，学習用データを整備した．表 1 に適用したルールの一部を示す．

表 1　学習データ収集に用いたルール

属性	人手によるルール
男性	男性です, 男です
女性	女性です, 女です
10 代	高校生です, 高 1 してます, 高 2 してます，中学生です
20 代	20 代，大学 4 年生です, 社会人 2 年目です
30 代	30 代，三十代
40 代	40 代，四十代
50 代	50 代，五十代
60 代	60 代，六十代，還暦，定年退職
全年代	*年生まれ

表 2　構築したユーザ属性推定モデルの評価

正解データ	属性	Precision	Precision 誤差許容	Recall
Facebook	性別	0.84	-	0.42
	年代	0.50	0.80	0.19
芸能人	性別	0.82	-	0.66
	年代	0.51	0.91	0.19

上記のルールを用いて，年代推定については年代ごとに 6000 件の学習用データ，性別推定については性別ごとに 1500 件の学習用データを収集した．

3.1.3　性能

収集した学習用データと正解ラベルに SVM を適用し，ユーザ属性ごとのモデルを構築した．これらを別途用意したテストデータにより評価を行った．テストデータとしては下記の 2 種類を用いた．

- Facebook ユーザ
  Twitter のプロフィール欄に自身の Facebook アカウントの URL を記述してるユーザを抽出し，各ユーザの Facebook の基本情報から，ユーザの属性情報（性別，年齢）を抽出し，それを正解ラベルとして用いた．用意したユーザは全部で 200 ユーザである．
- 芸能人ユーザ
  Twitter の芸能人アカウントを抽出した後，Wikipedia から芸能人の属性情報（性別、年齢）を抽出し，それを正解ラベルとして用いた．用意したユーザは全部で 200 ユーザである．

上記について評価した結果は表 2 に示す．下記のように性別については約 0.8 の Precision，約 0.4 の Recall を得ることができた．一方年代については約 0.5 の Precision，約 0.2 程度の Recall を得ることができた．年代判別の Precision は低いようにも考えられるが，6 クラス分類問題なので，ランダムに選択した場合の約 3 倍である．また，1 世代ずれを許容した場合の Precision は 0.80 であり，性別と同程度の精度である．

表 3　獲得されたユーザ属性の例と簡易な分類

属性種類	属性名
興味・関心	サッカー, 野球，アニメ（女性），アニメ（男性），ゲーム，テーマパーク，創作（小説，絵，歌）
ファン	アイドル，ジャニーズ, 女性声優 男性声優，ミュージシャン（J-POP，K-POP）
政治思想	自民党支持，民進党支持
職業	研究者，トレーダー，エンジニア
地域高校	静岡県, 栃木県，大阪府，沖縄県
地域大学	東京都, 九州，中部，近畿

3.2　Twitter 上のインタラクションデータを用いたユーザ属性構築

3.2.1　コミュニティ抽出手法を用いたユーザ属性構築手法

既存のユーザ属性によらない新たなユーザ属性の構築を試みる．本稿では，ユーザ属性構築に丸井，鳥海らによるインタラクションベースの手法を適用する [6], [9]. この手法は，「Twitter 上で相互にメンション [*1] しあっているユーザは類似した興味・関心を持っている」という仮定に基づき，Twitter ユーザの相互メンションネットワークを構築した後, そこから抽出したコミュニティをユーザ属性として用いる手法である．

具体的には下記のような手順により，コミュニティを抽出した後，ユーザのプロフィール文を用いて各コミュニティに属性名を付与する．

- Twitter データから，ユーザの相互メンション関係を抽出し，ネットワークを構築する．ここでは当社内に蓄積された 10%サンプリングデータ 1 ヶ月分を用いる．
- コミュニティ抽出手法の一つである Louvain 法を適用し，コミュニティを抽出する．
- コミュニティごとに，含まれるユーザのプロフィール文を収集し，コミュニティ文書を作成する．
- 全コミュニティ文書について，TF-IDF を適用し，コミュニティごとの特徴語 100 語を抽出する．
- コミュニティの特徴語上位 50 語を用いて，Wikipedia を検索し，検索結果上位 3 件の記事名をラベルとして抽出する．
- 得られた特徴語群とラベルを参考に，人手によりユーザ属性名を付与する．

3.2.2　獲得されたユーザ属性

適用手法により獲得されるユーザ属性の例を表 3 に示す．実際には，興味関心や地域，政治思想，職業など多様な軸のコミュニティが得られていることがわかる．

4.　実ビジネスでの活用

4.1　ユーザ属性別の集計によるソーシャルリスニング

得られたユーザ属性を用いて，様々なソーシャルメディ

[*1] 投稿内で相手のユーザ名を言及すること．言及されたことは相手に通知される．

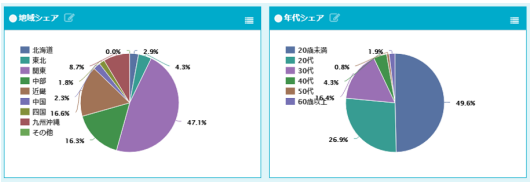

図 2 ユーザ属性を用いた実サービスの例

表 4 広島県の高校コミュニティにおける特徴語

単語	説明
jacklion	大阪にあるライブホール
誠之館	広島県にある高校名
福山駅	広島県にある駅名
オドループ	曲名
カラオケキャス	カラオケをライブ配信すること

アデータを集計することで，様々な知見を得ることができる．例えば，既存ユーザ属性を用いて集計することで，既存の社会調査やマーケティングの知見と比較することが可能となる．当社では，図 2 のように特定のキーワードで収集したツイート集合について既存ユーザ属性ごとに集計する機能をマーケティングツールとして提供している．また，特定のアカウントのフォロワーについて，どのようなユーザ属性の分布になっているかを可視化する機能を提供している．

このように特定の話題に関するツイート群や特定のユーザのフォロワーなど，様々なユーザ群を属性ごとに集計することでマーケティングに有用な知見を得ることができる．

4.2 特定ユーザ属性における話題

得られたユーザ属性に含まれるユーザ群の投稿を収集し，分析することで，そのユーザ属性における話題を抽出することができる．例えば，広島県の高校コミュニティに特徴的な語を抽出すると、表 4 のような語が得られる．広島県内の地名が得られると同時に，大阪にあるライブホールや流行の曲名，最近の若者文化が垣間見える．

このようにユーザ属性ごとに投稿を収集し分析することで，特定のユーザ属性の興味やライフスタイルの手掛かりを得ることができる．

5. おわりに

本稿では，ソーシャルメディアユーザ属性を推定し，それを実サービスとして活用している事例について紹介した．実際に大規模に適用する際の学習データやテストデータの収集方法を紹介し，さらに実際にサービスとして活用している事例を紹介した．

今回適用している手法は，ごく一般的なものを用いており，それ自体に新規性はない．他方，それらを適切に組み合わせることで，ビジネス上の課題を解決することに成功している．このように学術的な技術の新規性や高度さに対して，ビジネス上得られる効用が必ずしも相関しているわけではない．

今後，著者としては，ビジネス上の課題を適切に解決した AI 技術の適用事例を集積し，それらを分析することで，AI 技術の適切かつ効率的な活用方法についての枠組み構築を模索していきたい．

参考文献

[1] Backstrom, L., Sun, E. and Marlow, C.: Find Me If You Can: Improving Geographical Prediction with Social and Spatial Proximity, *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, ACM Press, pp. 61–70 (2010).
[2] Burger, J. D., Henderson, J., Kim, G. and Zarrella, G.: Discriminating Gender on Twitter, *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, Association for Computational Linguistics, pp. 1301–1309 (2011).
[3] Cheng, Z., Caverlee, J. and Lee, K.: You Are Where You Tweet : A Content-based Approach to Geo-Locating Twitter Users, *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, CIKM '10, ACM Press, pp. 759–768 (2010).
[4] Hecht, B., Hong, L., Suh, B. and Chi, E. H.: Tweets from Justin Bieber's Heart: the Dynamics of the Location Field in User Profiles., *Proceedings of the 2011 Annual Conference on Human factors in Computing Systems*, CHI '11, ACM Press, pp. 237–246 (2011).
[5] Ito, J., Hoshide, T., Toda, H., Uchiyama, T. and Nishida, K.: What is he/she like?: Estimating Twitter user attributes from contents and social neighbors, *2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2013)*, pp. 1448–1450 (2013).
[6] Marui, J., Nori, N., Sakaki, T. and Mori, J.: *Empirical Study of Conversational Community Using Linguistic Expression and Profile Information*, pp. 286–298, Springer International Publishing (2014).
[7] 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, ACM Press, pp. 37–44 (2010).
[8] 池田和史，　服部元，松本一則，小野智弘，東野輝夫：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，情報処理学会論文誌コンシューマ・デバイス＆システム（CDS），Vol. 2, No. 1, pp. 82–93 (2012).
[9] 鳥海不二夫，　榊剛史：バースト現象におけるトピック分析，情報処理学会論文誌，Vol. 58, No. 6, pp. 1287–1299 (2017).