1M4-5

投稿時間のクラスター分析によるTwitterユーザの年齢層推定

Age Estimation on Twitter User by cluster anlysis of tweet activity pattern

伊集竜之 *1 Tatsuyuki Iju　遠藤聡志 *2 Satoshi Endo　山田孝治 *2 Koji Yamada　當間愛晃 *2 Naruaki Toma　赤嶺有平 *2 Yuhei Akamine

*1琉球大学大学院理工学研究科情報工学専攻
Graduate School of Information Engineering, University of The Ryukyu

*2琉球大学工学部情報工学科
School of Information Engineering, University of The Ryukyus

Twitter is a social networking service, which is widely used worldwide and has enormous number of users. Thus there are various researches on estimating latent user attributes, including gender, age, and so on. The results of estimation is applied for some purpose such as advertising, or utilizing Twitter as social sensor. In this study, we focus on “ age ”, which is one of the most basic attributes. In order to estimate the age of Twitter user, we employ “ tweeting activity ” as user's lifestyle and words vector of tweet text. We propose an estimating method considering“ tweeting activity model ”that is constructed by clustering users on their tweeting activity and examine effectiveness of our method by computer experiments.

1. はじめに

現在、企業などにおいて膨大に蓄積されたデータを解析し、ビジネスに活用しようという動きが活発化している。解析対象の１つとして大きく注目を浴びているのがマイクロブログである。代表的なマイクロブログである Twitter は、2015 年には日本において、2230 万人のユーザを抱えるまでに成長すると予測されている [1]。Twitter における特徴的な機能として、特定のユーザの Tweet を追跡する follow、自分とは異なるユーザの Tweet を発信する Retweet、Tweet の内容が特定の話題に関する内容であることを明示する hashtag などがある。これらの機能はユーザが他の不特定多数のユーザへ情報を発信することを容易にし、Twitter 上における情報の即時性や、拡散性を高めている。この Twitter の特性に着目し、Twitter をソーシャルセンサー [2][3] として活用する取り組みが行われているが、そのような場合、ユーザの年齢、性別、社会的な役割などの基本的な属性の違いを考慮することがより効果的である。しかし、Twitter ではユーザは自身の情報を自身の裁量で公開することができ、狙った属性情報を得ることは難しい場合が多い。

本研究ではユーザの発信する情報から、属性を推定する手法を提案する。推定する属性としては「年齢層」に着目する。年齢層は最も基本的な属性の１つであり、他の属性の推定に有効な情報の１つである。

2. 関連研究

Burger[4] らはユーザの Tweet のテキスト、スクリーンネーム、自己紹介文を推定に利用し、Twitter ユーザの性別の推定を行った。Pennacchiotti[5] らは Burger らと同様に、Tweet のテキスト、スクリーンネーム、自己紹介文に加え、follow 関係の情報を活用し、ユーザの政治思想、特定のビジネスへの興味の有無、属する民族の推定を行った。Rao[6] らは Tweet テキスト中の所有格表現の後に続くワードに着目し、ユーザの性別、年齢層、出身地、政治思想の推定を行った。

連絡先: 伊集竜之 琉球大学 k148582@ie.u-ryukyu.ac.jp

以上の様に Twitter ユーザの属性推定については様々な検討がなされているが、そのほとんどが過去の Tweet の内容や自己紹介文などのユーザが発信するテキスト情報、または follow 関係に着目している。しかし、年齢層推定の場合、これらとは別の情報として「ユーザのライフスタイル」の活用も有効であると考えられる。ライフスタイルは社会的な役割に強く影響を受けるが、一般的に社会的な役割の範囲は年齢によって定まる場合が多く、年齢によってライフスタイルをある程度推定することが可能である。ライフスタイル情報を活用した推定手法の検討は盛んではなく、活用方法は確立されていない。したがって、本研究ではユーザのライフスタイル情報を考慮することの有効性を検証する。

3. 年齢層

年齢層の定義について述べる。本研究では、30 歳以上と 30 歳未満の 2 つの年齢層を定義する。中小企業庁による中小企業白書の生産年齢人口に関する統計 [7] では、15 歳から 30 歳未満を若年層として区別しており、一部の国家公務員試験では受験資格の１つとして「満 30 歳未満」を定めている [8]。このように、30 歳という年齢は社会的なキャリアにおける１つの区切りであると広く認知されている。したがって、30 歳以上であるか、未満であるかを推定することはユーザのモデリングやユーザの理解に役立つと考えられる。

4. ライフスタイルに着目した年齢層推定手法

ライフスタイルの情報は Tweet のテキストから得られるような明示的なものではなく、暗黙的なものである場合がほとんどであると考えられる。したがって、本研究ではライフスタイル情報の捉え方について様々な検討を行った。その結果として以下の「投稿アクティビティベクトル」の設計を行う。また、「投稿アクティビティベクトル」を用いて教師あり学習による分類器を構築し、分類器の性能評価によって投稿アクティビティベクトルの有効性を検証する。

4.1 投稿アクティビティベクトル

Twitter の持つ情報の即時性や投稿の容易さなどから、Twitter ユーザのライフスタイルに現れる特徴はユーザの Tweet 投稿アクティビティからある程度捉えることができる可能性が高い。そこで本研究では「投稿アクティビティベクトル」を設計し、推定における有効性を検証する。投稿アクティビティベクトルの具体的な構築方法は以下の通りである。

ステップ 1: $T_{u_i,wd}$ 及び $T_{u_i,hd}$ を計算する。$T_{u_i,wd}$ はユーザ u_i が月曜から金曜日において時間 t に発信した Tweet 数の割合を表しており、$T_{u_i,hd}$ は土曜日及び日曜日において時間 t に発信した Tweet 数の割合を表している。それぞれ以下の様に計算される。

$$T_{u_i,wd}(t), T_{u_i,hd} = \frac{\sum_{n=1}^{d} T_{u_i,n}(t)}{\sum_{n=1}^{d}\sum_{t=0}^{23} T_{u_i,n}(t)} \quad (1)$$

$$n = 1, \cdots, d \quad (2)$$

$$t = 0, \cdots, 23 \quad (3)$$

ステップ 2: $T_{u_i,wd}$ 及び $T_{u_i,hd}$ 結合し、最終的なアクティビティベクトル $T_{u_i} = \{T_{u_i,wd}, T_{u_i,hd}\}$ を作成する。

作成された T_{u_i} はユーザ u_i の全 Tweet における、平日 (月曜から金曜) および週末 (土曜、日曜) の 1 時間毎の Tweet 投稿数の割合を表している。すなわち、T_{u_i} の構成は平日に着目した投稿アクティビティと週末に着目した投稿アクティビティの 2 つからなるが、これにはユーザの平時におけるライフスタイルと余暇におけるライフスタイルの違いを汲み取り易くするという狙いが存在する。

最終的なアクティビティベクトルは以下の様な 48 次元のベクトルとなる。

$$T_{u_i} = \{T_{u_i,0}, T_{u_i,1}, \cdots, T_{u_i,47}\} \quad (4)$$

4.2 単語ベクトル

従来手法では、Tweet の内容に出現する特徴的な単語を素性とした「単語ベクトル」が最も広く用いられている。これはユーザの属性を推定しようとする場合、Tweet の内容が最も大きな情報源となるためである。本研究では、投稿アクティビティベクトルと単語ベクトルを組み合わせることにより、単語ベクトルでは捉えきれない情報を投稿アクティビティベクトルが補完するかどうかについても検証する。単語ベクトルの構築方法は以下の通りである。

ステップ 1: ユーザ集合 $u = \{u_1, u_2, \cdots, u_n\}$ に含まれるユーザから Tweet を 200 件抽出し、それらの Tweet 中に出現する単語を記録し、得られた単語の中から、χ^2 値で上位 16000 件の単語を選び出す。Tweet の件数を 200 件に制限した理由は、ユーザ毎の Tweet 量の偏りや 1Tweet 内の文章量の不均一を考慮さを考慮したためである。また、χ^2 値で上位 16000 件の単語を選び出す理由は、各年齢層クラスに特徴的である推定に有効な単語のみを利用するためである。

ステップ 2: ステップ 1 で選び出された単語を用いて、ユーザ u_i を表すベクトル W_{u_i} を構築する。単語ベクトルの値は u_i の 200 件の Tweet 中での単語の出現割合である。

最終的な単語ベクトルは以下の様な 16000 次元のベクトルとなる。

$$W_{u_i} = \{W_{u_i,0}, W_{u_i,1}, \cdots, W_{u_i,15999}\} \quad (5)$$

4.3 実験

ユーザのライフスタイルに着目した、投稿アクティビティベクトルの年齢層推定における有効性について検証する。

4.3.1 実験データ

実験データは Yahoo!ブログまたは FC2 ブログのアカウントを持つ Twitter ユーザから収集した。収集方法の詳細は以下の通りである。

ステップ 1: Twitter の SearchAPI を用い、アカウントに Yahoo!ブログまたは FC2 ブログの自分のブログへのリンクを登録しているユーザを収集する。

ステップ 2: リンク先ブログのユーザプロフィール情報にユーザの生年情報が登録されている場合は生年情報を取得し、合わせてそのユーザの Tweet も取得する。ステップ 1 で収集したユーザ全てについてこの処理を繰り返し、ユーザデータベースを構築する。

ステップ 3: 年齢が 13 歳未満または 80 歳以上のユーザをデータベースから取り除く。これは、極端に年齢の低いユーザや高いユーザは年齢を詐称している可能性が高く、そのようなユーザを取り除くためである。13 歳が下限値となっている根拠は Twitter の利用規約により、13 歳未満の者は基本的に Twitter の利用を認められていないためである。

以上の処理の後、最終的に実験データとして使用できるユーザとして、30 歳未満で 2542 人、30 歳以上で 1459 人を得た。

4.3.2 実験方法

以下の 3 つの場合それぞれについて SVM で分類器を構築し、5 分割交差検定を用いて性能を評価する。

1. 投稿アクティビティベクトルを用いる場合
2. 単語ベクトルを用いる場合
3. 投稿アクティビティベクトル及び単語ベクトルを用いる場合

評価尺度には精度、適合率、再現率、F 値を利用する。なお、SVM のパラメータについては、学習データ及びテストデータとは別に用意した各年齢層それぞれ 215 人のユーザに対してグリッドサーチを行い、kernel を linear、C の値を 100 と決定した。

4.3.3 実験結果

表 1 に実験結果を示す。手法 1 は投稿アクティビティベクトルを用いた場合、手法 2 は単語ベクトルを用いた場合、手法 3 は単語ベクトル及び投稿アクティビティベクトルを用いた場合の結果をそれぞれ示している。投稿アクティビティベクトルのみを用いた場合については、適合率、再現率、F 値、精度それぞれで約 0.6 ポイントを得た。テストデータのユーザは各年齢層で同数であるので、精度のベースラインは 0.5 ポイントとすれば、これを 0.1 ポイント上回る 0.6 ポイントという結果は、Tweet の投稿アクティビティのみでユーザの年齢層をある程度推定することが可能であることを示している。投稿アクティビティベクトルはライフスタイルという暗黙的な情報を捉えたものである。したがって、投稿アクティビティベクトルと単語ベクトルを組み合わせることによって、単語ベクトルでは捉えきれない特徴を投稿アクティビティベクトルが補完し、単語ベクトル単体の場合よりも精度が向上することが期待される。しかし、結果を確認すると精度、F 値は単語ベクトル単体の場合とほぼ同じ値となっている。よって、投稿アクティビティベクトルと単語ベクトルの単純な組み合わせでは期待していたような分類器の性能向上は得られなかった。しかしながら、投稿アクティビティベクトルの有効性は実験結果より示されている

表 1: 実験結果

	手法 1		手法 2		手法 3	
	30 歳未満	30 歳以上	30 歳未満	30 歳以上	30 歳未満	30 歳以上
適合率	0.60	0.60	0.767	0.740	0.749	0.759
再現率	0.584	0.619	0.726	0.778	0.764	0.742
F 値	0.593	0.609	0.746	0.758	0.756	0.750
精度	0.60		0.752		0.753	

ため、より発展的な推定手法による精度向上の余地は残されていると考えられる。

5. 投稿アクティビティモデルによる年齢層推定手法

前章の実験結果から、単語ベクトルと投稿アクティビティベクトルの単純な組み合わせでは分類器の性能向上は得られなかった。一般的に同じ年齢層であっても様々なライフスタイルが存在するため、単純な組み合わせでは単語ベクトルが捉えきれない特徴を投稿アクティビティベクトルが捉えるまでには至らなかったのではないかと考えられる。したがって、発展的な推定手法として、各年齢層で特徴的な投稿アクティビティのモデルを数個構築し、ユーザの年齢層の推定の際にはユーザの投稿アクティビティとの類似度を考慮する手法の提案を行う。

5.1 投稿アクティビティモデル

投稿アクティビティモデルの構築法について述べる。

ステップ 1: $u = \{u_1, u_2, \cdots, u_n\}$ に含まれるユーザ u_i の時間当たりの Tweet 投稿割合からアクティビティベクトル T_{u_i} を構築する。この処理を u 全体に対して行い、$T_u = \{T_{u_1}, T_{u_2}, \cdots, T_{u_n}\}$ を構築する。

ステップ 2: 作成された T_u を用いて、各年齢層毎に K-means 法を利用してクラスタリングを行い、でき上がった各クラスタの中心を u の投稿アクティビティモデルとする。K-means 法における K の値すなわちクラス数は、検証実験で利用した SVM のパラメータを決定する際に用いたデータと同じデータ上で、$K = 2 \sim 5$ の範囲でクラスタリングを行い、最も評価値の高かった $K = 2$ に決定した。評価値の算出には Silhouette Coefficient[9] を使用した。

5.2 単語モデル

検証実験の結果より、単語ベクトルは投稿アクティビティベクトルと比較して有効性が高いことが分かった。したがって、年齢層に特徴的な単語をモデル化した単語モデルを構築し、年齢層推定に活用する。単語モデルの構築方法は、4.1.1 で述べた方法と同様の方法で、ユーザ集合 $u = \{u_1, u_2, \cdots, u_n\}$ から $W_u = \{W_{u_1}, W_{u_2}, \cdots, W_{u_n}\}$ を構築し、これを SVM への入力として単語モデルを構築する。

5.3 年齢層推定処理

提案手法における年齢層推定処理について述べる。本処理では、投稿アクティビティモデルが算出する類似度と単語モデルが算出する類似度を基に、対象とするユーザが属する年齢層クラスを推定する。推定プロセスは以下のステップから成る。

ステップ 1: 単語モデルを用いて対象とするユーザ s_i が、クラス c に属する確率 $P(s_i, c)$ を算出する。$P(s_i, c)$ の算出には SVM の実装である LIBSVM が持つ機能である、Predict-Probability を利用する。

ステップ 2: Tweet 投稿パターンモデルで生成された、クラス c に含まれるクラスタ m の中心とユーザ s_i の Tweet 投稿アクティビティベクトルの類似度 $Sim(s_i, c, m)$ を算出する。類似度の尺度には χ^2kernel を使用する。また、$Sim(s_i, c, m)$ は $Sim(\hat{s_i}, c, m)$ として以下の様に正規化される。

$$Sim(\hat{s_i}, c, m) = \frac{Sim(s_i, c, m)}{\sum_{n \in \{o30, u30\}} \sum_{k=1}^{l} Sim(s_i, n, k)} \quad (6)$$

$o30$、$u30$ はそれぞれ 30 歳以上、30 歳未満の年齢層クラスを表している。l は年齢クラス別のクラスタ数の合計数である。

ステップ 3: $P(s_i, c)$ と $Sim(s_i, c, m)$ を統合した統合類似度 $I(s_i, c, m)$ を算出し、

$$I(s_i, c, m) = P(s_i, c) \times Sim(s_i, c, m) \quad (7)$$

最も高い値を持つ $I(s_i, c, m)$ が属する年齢層クラスをユーザ s_i が属するクラスと推定する。

5.4 提案手法の検証実験

実験では、投稿アクティビティモデルを利用した提案手法の有効性を検証する。実験に用いるデータは前章の実験で用いたデータと同一である。前章と同様に提案手法を 5 分割交差検定で評価し、評価尺度には精度、適合率、再現率、F 値を利用した。表 2 は、その結果である。

実験の結果として、分類器の性能向上はさほど得られなかった。投稿アクティビティモデルとの類似度比較によって様々なライフスタイルの存在に対応し、推定精度の向上を図ったが、有効な結果が得られたとは言い難い。本手法ではユーザの投稿アクティビティモデルの数を各年齢層で 2 つとしていたため、様々なライフスタイルの存在に十分に対応しきれなかったことも一因と考えられる。

表 2: 提案手法の検証実験結果

	30 歳未満	30 歳以上
適合率	0.759	0.756
再現率	0.755	0.757
F 値	0.756	0.757
精度	0.757	

6. まとめ

本研究では、Twitter ユーザの年齢層推定を行った。年齢層を推定する情報としてユーザのライフスタイルに着目し、ライフスタイルを表す情報として投稿アクティビティベクトル設計し、分類器を用いた推定実験を行った。その結果投稿アクティ

ビティベクトルは推定に有効な情報が含まれることが分かった。また、投稿アクティビティモデルを構築し、ユーザの様々なライフスタイルの存在に対応するというアイディアの下、モデルとの類似度によってユーザの年齢層を推定する手法を提案したが、推定精度の向上はさほど得られなかった。今後、より詳細なライフスタイルのモデル化によって精度向上を図る手法を検討していく。

参考文献

[1] Asia-Pacific Grabs Largest Twitter User Share Worldwide
http://www.emarketer.com/Article/Asia-Pacific-Grabs-Largest-Twitter-User-Share-Worldwide/1010905

[2] インフルくん http://mednlp.jp/influ/

[3] 那須野薫, 松尾豊:Twitter における候補者の情報拡散に着目した国政選挙当選者予測, 人工知能学会, 2014.

[4] John D. Burger, John Henderson, George Kim and Guido Zarrella. Discriminating Gender on Twitter. In Proceedings of Conference on Emprical Methods in Natural Language Processing, pages 1301-1309, 2011

[5] Marco Pennacchiotti and Ana-Maria Popescu. Democrats, Republicans and Starbucks Afficionados: User Classification in Twitter. KDD, 2011

[6] Delip Rao, David Yarowsky, Abhishek Shreevats and Manaswi Gupta. Classifying Latent User Attributes in Twitter.

[7] 中小企業庁編. 中小企業白書. 平成 22 年版

[8] 国家公務員採用情報. 内閣官房府 http://www.cas.go.jp/jp/gaiyou/jimu/jinjikyoku/recruit/howto/senmon.html

[9] PeterJ. Rousseeuw(1987).”Silhouttes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis”. Computational and Applied Mathematics 20: 53-65.