第 24 回インテリジェント・システム・シンポジウム [2014.9.18－19]

FAN-14-027

Twitter ユーザの年齢層推定のための有効な素性の検討

Effective Features For Estimating Age of Twitter Users

○伊集　竜之（琉球大学）
遠藤　聡志（琉球大学）　山田　孝治（琉球大学）　當間　愛晃（琉球大学）　赤嶺　有平（琉球大学）

Tatsuyuki IJU, k148582@ie.u-ryukyu.ac.jp
Satoshi ENDO　Koji YAMADA　Naruaki TOMA　Yuhei AKAMINE
University of Ryukyu

Twitter has become an important social networking service , which is widely used worldwide and has enormous number of users. Thus the ability to estimate latent user attributes, including gender, age, and so on, from contents of tweet and other information (e.g. friends, followers) has important applications in advertising and recommendation. We focus on " age ", which is one of the most basic attributes, and we propose SVM-based algorithm for classifying the age of uncharacterized Twitter user. In this study, we have defined age as four classes(13-19, 20-35, 36-60, 61-). We also carry out several experiments about five features that are extracted from a text of Tweet and evaluate the effectiveness of each features.

Key Words: Twitter, Feature, age, SVM

1　はじめに

Twitter は最大 140 字の短文を投稿できる SNS である。Twitter の持つ情報の拡散力や即時性、膨大なユーザ数という特徴から、Twitter ユーザーの人物像は、マーケティングやユーザー推薦などに利用できる。しかし、パーソナルデータを提供するユーザー数は全体から比較すれば、極小数である。 したがって、ユーザーの発信する情報からユーザーの人物像を推測する技術が求められる。本研究では、人物像のひとつとして、「年齢層」に着目し、ユーザーの年齢層を推定する手法を提案する。年齢層は、 人物像を推定する上で最も基本的な情報の一つである。

Twitter ユーザの人物像推定の研究としては、長浜ら [1] の研究がある。長浜らはユーザのツイートの内容からそのユーザの性別の推定に、4 つの素性を利用する手法を提案した。性別が予め分かっているユーザのツイートに対して形態素解析を行い、「単語」、「特定の条件を満たす単語」、「品詞」、「品詞並び」のそれぞれをもとに、SVM を利用して分類器を作成し、その精度を確認した。なお、素性値として「品詞」と「品詞並び」は出現割合を利用し、その他は χ 二乗値を利用した。χ 二乗値は統計指標の一つであり、要素のカテゴリにおける特徴量を表す。実験の結果、平均としてそれぞれ、83.7%、81.3%、79.7%の精度を得た。そこで、本研究は、年齢層を対象とする。

Web テキストを用いた年齢層推定の研究として、泉ら [2] は投稿記事のみで blog 著者の年齢層を 10 代 (13-17 歳),20 代 (23-27 歳),30 代 (33 歳-42 歳), その他の 4 つのクラスに分類する手法を提案し た。著者の年代が予め分かっている blog 記事に対し、形態 素解析を行い、全形態素に対してエントロピーと相対出現 頻度を算出することで推定に有用な 2 グラム単語を抽出、これらの素性をもとに単純ベイズ分類器を作成し、その精度を確認した。なお、精度指標として F 値を利用した。その 結果、10 代に対して 0.86、20 代に対して 0.57、30 代に対 して 0.96、全体で 0.71 の精度を得た。本研究とは、年齢層を推定するという点では共通だが、blog 著者を対象としている点で異なる。

2　年齢層

2.1　本研究における年齢層定義

E.H.Erikson[3, 4, 5] によれば、人間の心理的または社会的発達の段階は 8 つに区分できる。その内訳は、乳児期 (0 歳 - 2 歳)、幼児前期 (2 歳 - 4 歳)、幼児後期 (4 歳 - 5 歳)、児童期 (5 歳 - 12 歳)、青年期 (13 歳 - 19 歳)、成人期 (就職 - 結婚)、壮年期 (子育て - 退職)、老年期 (退職後) である。成人期以降については、範囲が曖昧であるが、範囲を具体化する際の目安が与えられており、それらは括弧内の記述の通りである。本研究では、Erikson の定義をもとにし、同じ発達段階にあるユーザの Tweet には共通な特徴が現れると仮定して表 1 のような新たな区分を定義し、利用する。

それぞれ、青年期、成人期、壮年期、老年期に対応している。13 歳未満のユーザを含まない理由は、Twitter 社が規約により、利用を禁じているためである。また、B、C、D 層のそれぞれの年齢範囲は、日本人の平均値 [6] を基準に定めた。

Table 1 本研究における年齢層定義

A 層	B 層	C 層	D 層
13 歳 - 19 歳	20 歳 - 35 歳	36 歳 - 60 歳	61 歳以降

3　素性

本研究では、Tweet のテキストに由来する素性として、長浜らの提案した素性 (単語、品詞、品詞並び、特定条件を満たす単語) に加え、新たに「共起語」を採用する。また、ユーザーアカウントに属する情報に由来する素性として、ユーザーのフレンド数、フォロワー数、フレンド数とフォロワー数の比率を採用する。

3.1 共起語

テキストの書き手の属性を推定する場合、テキスト中に出現する形態素を単体で見るよりも、形態素の組み合わせを見る方が得られる情報が大きいと考えられる。したがって、n グラムや共起語を素性として利用することが有効であると考えられるが、ユーザーによって文体にかなりの違いがある Tweet では、n グラムが持つ形態素の「並び」という情報はノイズとなる可能性が高い。したがって、本研究では、「共起語」を新たな素性として採用し、その有効性を確認する。「共起語」は、テキスト中に長さ 3 のウィンドウを設け、そのウィンドウ内に出現する単語の全てのペアを「共起語」として用いる。さらに、ウィンドウをテキスト中の文頭から文末までスライドさせることによって、テキスト中の全共起語を抽出する。図 1 に「今日は数学のテストだ」という内容のツイートから共起語の抽出を行う場合の例を示す。

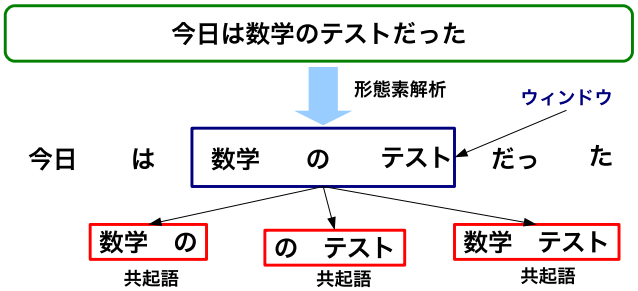

Fig.1 共起語抽出の例

この例では、(数学, の)、(の, テスト)、(数学, テス ト) という共起語が抽出されることが示されている。ここで、(数学, テスト) という共起語に着目する。テストという単語が単体で与えられた場合を仮定すると、このアカウントは学生であることが推定されるが、同時に数学という単語が与えられたならば、中学生以上の学生であると推定でき、より具体的な予想を行うことができる。

3.2 ユーザーアカウント情報素性

Twitter ユーザの人物像を推定する場合、そのユーザーの投稿した Tweet の内容は最も大きな手がかりとなるが、ユーザーのアカウントそのものに属する情報 (以下では、単純にユーザーアカウント情報と呼ぶ) もまた、有用な手がかりであると言える。例として、ユーザのスクリーンネーム、自己紹介欄のテキスト、アバターイメージ、フレンド · フォロワー情報などが挙げられる。ユーザーアカウント情報は、Tweet と同様にユーザーの趣味嗜好やバックグラウンドを反映すると考えられるが、Tweet と比較して、社会のトレンドの変化による影響を受けにくく、ノイズとなるような情報が混じりにくいと考えられる。本研究では、ユーザーのフレンド · フォロワー情報に着目し、予備実験として、ユーザーのフレンド数、フォロワー数、フレンド数とフォロワー数の比率を素性として採用し、推定実験を行う。

4 年齢層推定実験

4.1 実験概要

本研究では、SVM を分類器として利用した教師あり学習によって、分類モデルを構築し、ユーザーの年齢層推定を行う。以下は、実験の流れである。

1. A、B、C、D の年齢層ラベルを付与されたユーザーの個々に対し、Tweet を形態素解析器 MeCab を用いて形態素解析し、分解された形態素を素性に変換する。素性を用いて、それぞれのユーザーを素性ベクトルで表現する。素性値は、素性が品詞の場合はその出現割合、その他については、要素が出現した場合を 1、出現しない場合を 0 とする。

2. 素性ベクトルを利用して SVM を用いて学習させ、分類実験を行う。本研究では、SVM を多クラス分類へ対応させる手法として、ペアワイズ法 [7] を利用した。また、学習と精度評価には、5 分割交差検定を利用した。

4.2 実験データ

実験データとして、生年が判明しているユーザの 2013 年 6 月から 12 月にかけての Tweet を収集した。ユーザーの生年の確認は、Twitter の自己紹介欄の記述や、ユーザの持つ Blog のプロフィールの生年月日の入力を基に人手で行った。ユーザの人数構成は A、B、C、D の各年齢層で 50 人ずつである。

4.3 各年齢層に特徴的な形態素

各層別に χ 二乗値で上位 20 件の形態素を表 2 に示す。A 層、D 層では他の層と比較して、顕著な差異が現れている。A 層の特徴は以下の通りである。

1. 「まじ」などいった、口語的な表現が多い。

2. 「「」、「ー」などの記号が上位にある。

3. 「おやすみ」、「ただいま」といった挨拶の際に用いられる表現が上位にある。

A 層のユーザーの Tweet を確認してみると、友人など他のユーザとのコミュニケーションツールとして利用する傾向が強く、1 と 3 の結果はそれを反映したものであると考えられる。また、「ー」などの記号は感情を表す表現によく付随されるであり、2 の結果もこの傾向を捉えた結果であると考えられる。「部活」という「部活動」を表す表現が上位にあるが、この表現は中高生によく用いられる表現であり、A 層が中高生を中心としたユーザーで構成さているという特徴を強く反映した結果であると言える。

D 層については、「自民党」、「自民」、「維新」などの政党を表す名詞が並んでいる。D 層は 65 歳以上のユーザで構成された層であり、高齢者の多い D 層のユーザは政治に対して強い関心を持っていると推測される。

B、C 層については、一見して関連性を指摘するのが難しいワードが多く並んでおり、表の内容を基にユーザの特徴を推測することは困難である。したがって、B、C 層のユーザーの Tweet の内容を確認したところ、両層の特徴として、自己啓発やビジネスへ関心を持つ傾向にあるユーザーが多いことが分かった。これは、B、C 層が 20 代から 50 代の働き盛りの年齢のユーザを中心に構成されていることを反映した結果であると言える。また、この結果から、「努力」、「経験」、「ビジネス」などのワードは先ほど述べた「自己啓発」や「ビジネス」に関する Tweet 内に現れるワードという共通した背景を持つと推測できる。

4.4 各素性の有効性に関する考察

表 3 は、各素性別に、全ての要素を使用した場合の精度をまとめたものである。「単語」を素性とした場合の精度

Table 2 各層別の χ 二乗値で上位 20 件の形態素

rank	A 層	B 層	C 層	D 層
1	よう	努力	量	福島
2	おやすみ	入れる	楽しく	提案
3	」	トイレ	鳥	4 月
4	「	たった	とれ	禁止
5	あいつ	親	ずいぶん	夏
6	ー	見つけ	いやぁ	台風
7	なり	瞬間	いくら	三
8	さ	動画	自身	様
9	部活	ビジネス	万	元
10	まじ	頼ま	画像	迎え
11	そう	邪魔	変え	賃金
12	前	逃げ	足	警察
13	より	迷惑	経験	訴え
14	6	若干	など	薬
15	さん	思考	声	自民党
16	どう	思える	でしょ	自民
17	9	嫌わ	青い	育ち
18	ただいま	始める	集約	維新
19	7	噂	重視	演説
20	-	メルマガ	還元	法人

が最も高く 66%であったが、「品詞」を素性とした場合が突出して精度が低かった以外は、全体として大きな差は見られなかった。

Table 3 各素性別の全要素を使用した場合の精度

素性	平均要素数	平均精度
単語	28000	66%
品詞	12	56%
品詞並び	37445	64%
共起語	152651	64%
特定条件を満たす単語	27935	63%

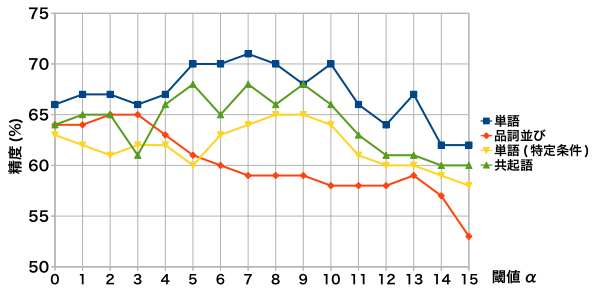

Fig.2 各素性別の精度の推移

図 2 は、「品詞」を除いた素性のうち、それぞれ年齢層別に、閾値 α 以上の χ 二乗値を持つ要素のみを選択し、利用した場合の精度の推移を表したものである。「品詞」を除いた理由としては、「品詞」は 12 種類しか存在せず、χ 二乗値を利用した要素選択がほとんど意味をなさないためである。

表 3、図 2 の結果を基に、実験に利用した素性についての考察を以下の様にまとめる。

1. χ 二乗値による要素選択を行わない場合、品詞を除き、いずれの素性も 6 割台の精度を得ており、有効性があると言える。しかし、精度自体はほぼ横並びであり、どの素性がより有効であるという結論を出すことは難しい。

2. χ 二乗値による要素選択を行った場合、いずれの素性についても 2 ～ 5%ほど精度が上昇するが、「品詞並び」に関しては、χ 二乗値の閾値を大きくすると、早々に精度が下降傾向に入るため、他の素性と比較して特徴量の大きな要素をあまり持たないと考えられる。したがって、仮に、χ 二乗値よりも優れた要素選択手法を用いたとしても、精度上昇の余地はさほど残されていないと思われる。

3. 1、2 より、単体で素性を用い、かつ、χ 二乗値などの手法を利用し、要素選択を行う場合、「単語」、「共起語」、「特定条件を満たす単語」の形態素をベースとした素性の方が、「品詞」、「品詞並び」の品詞をベースとした素性よりも有効であると思われる。

図 2 を参照すると、所々に精度のばらつきが現れているのが分かる。原因として B 層と C 層のユーザの特徴の類似性が考えられる。B 層と C 層のユーザーはともに働き盛りの年齢にあるユーザーを中心として構成されているという共通した背景を持つため、Tweet の内容がビジネスに関するものなど、似通った特徴を持つ傾向がある。したがって、B、C に関しては χ 二乗値による要素選択が上手く働かず、この 2 層のユーザの推定が不安定になり、結果的に全体として推定精度にばらつきが生じている可能性がある。今後、今回利用した年齢層定義とは異なる別の年齢層定義を利用した実験や、χ 二乗値以外の統計的要素選択手法を利用した実験を行い、結果を確認する必要がある。

4.5 素性を併用した手法の有効性についての考察

図 3 は要素数の推移を表した物である。閾値 α を増加させていくにつれ、いずれの素性についても要素数が減少していくことが分かる。ここで図 2 を参照すると、「品詞並び」を除き、いずれの素性についても、χ 二乗値の閾値 10 の点までは、ほぼ精度のピークを保っている。そこで、再び図 3 に戻り、χ 二乗値の閾値 10 の点での要素数に着目すると、当初他の素性と比較して圧倒的に要素数が多かった共起語を含め、いずれの素性についても、要素数が 1000 個ほどにまで大きく減少している。

図 4 は「単語」、「品詞並び」、「共起語」のそれぞれの素性ついて χ 二乗値 0 以上 の要素を利用した場合に、推定に成功したユーザの数をベン図で表現したものである。この結果より、以下の 3 点が確認できる。

- 3 つの素性で推定に成功した 108 人のユーザの内、43 人が A 層のユーザであり、A 層のユーザは比較的推定しやすいユーザである。
- B 層と C 層は共に最も少ない人数の 19 人であり、B 層と C 層のユーザーは比較的、推定に難しいユーザーである。
- 3 つの素性のそれぞれで、それぞれの素性でしか推定に成功出来ないユーザが存在する。

以上の結果から、素性を併用した手法の有効性についての考察は以下の通りである。

- χ二乗値による要素選択を行うことによって、ある程度精度を保ちつつ、学習に用いる素性ベクトルの次元数を大幅に削減でき、学習に費やす時間を短縮できる可能性がある。

- 「単語」、「共起語」、「品詞並び」のそれぞれについて、その素性でしか推定に成功することができないユーザーが数人おり、素性を併用した手法が有効である可能性が高い。また、そのような手法を用いる場合でも、χ二乗値を用いて要素選択を行うことで、学習に費やす時間を抑えることが可能であると思われる。

ユーザーアカウント情報素性を利用した場合の、推定実験の結果と考察は当日に報告する。

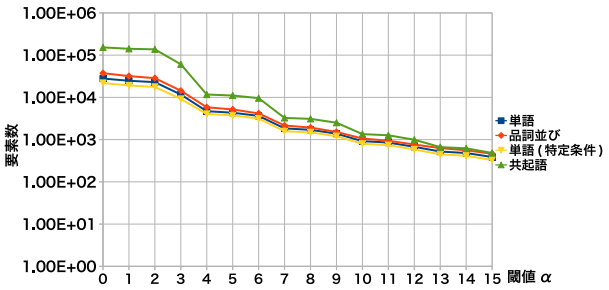

Fig.3 各素性別の要素数の推移

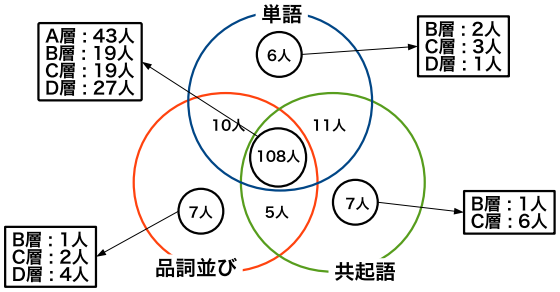

Fig.4 素性別の推定に成功したユーザー数

5 話題の変化が与える影響の分析のための実験

現実での応用を考える場合、あるイベントの影響がTwitter上で話題の変化として現れ、推定精度に影響を及ぼす場合があると仮定できる。本研究では、その仮定を検証し、話題の変化が推定精度にどのような影響を与えるかを分析するため、以下の様な4つのデータセットを定義し、実験を行う。

5.1 ベースデータセット

- 2013年7月 ～ 9月にかけて作成されたTweetのみで構成されたベースデータセットA

- 2013年10月 ～ 12月にかけて作成されたTweetのみで構成されたベースデータセットB

5.2 イベントデータセット

- 2014年1月 ～ 2014年3月にかけて作成されたTweetのみで構成されたイベントデータセットA

- 2014年5月 ～ 2014年7月にかけて作成されたTweetのみで構成されたイベントデータセットB

イベントデータセットAは、2014年2月に開催された冬期オリンピックに関するTweetが多く含まれると予想されるデータセットであり、データセットBは2014年6月に開催されたサッカーワールドカップに関するTweetが多く含まれると予想されるデータセットである。ベースデータセットA、Bは、特に世界的なイベントが起こらなかった期間に作成されたTweetを含んだデータセットである。

5.3 実験計画

1. ベースデータセットA + ベースデータセットBを利用した推定実験

2. イベントデータセットA + ベースデータセットAを利用した推定実験

3. イベントデータセットB + ベースデータセットAを利用した推定実験

2と3については、ベースデータセットAをベースデータセットBに変更した上で、再度、推定実験を行う。各実験での精度を比較し、話題の変化がどのように推定精度にどのような影響を与えるかを分析する。結果については、当日報告する。

6 まとめ

本研究では、Tweetの内容に由来する素性として、「単語」、「品詞」、「品詞ならび」、「共起語」、「特定条件を満たす単語」の5つを採用し、それぞれの年齢層推定においての精度を確認した。その結果、「品詞」を除き、いずれの素性についても6割台の精度を得ることができ、年齢層推定に有効であることが確認できた。さらに、「品詞」を除いた4つの素性についてχ二乗値による素性選択を行い、それぞれの素性について、精度の上昇が得られることを確認した。また、ある程度の精度を保ちつつ、素性ベクトルの次元数を大幅に削減でき、学習に費やす時間を短縮できる可能性があることを発見した。「単語」、「共起語」、「品詞並び」について、それぞれでしか推定に成功することができないユーザーが存在することが分かり、素性を組み合わせた推定手法が有効である可能性が高いという結果が得られた。

今後は、素性を併用した推定手法の提案を行い、有効性を評価する。また、異なる年齢層定義での実験や、χ二乗値とは異なる要素選択手法を用いての実験を行う。

References

[1] 長浜祐貴, 遠藤聡志, 當間愛晃, 赤嶺有平, 山田孝治:「ツ イート解析による性別推定に有用な因子の検討」第 12 回情報科学技術フォーラム (FIT2013), 2013

[2] 泉雅貴, 三浦孝夫, 塩谷勇:「Blog 著者年代推定のための エントロピによる特徴語抽出」電子情報通信学会第 19 回データ工学ワークショップ, 2008

[3] Erikson,E.H.(1959).Identityand the life cycle.Psychol. lss. 1:1-71.

[4] Erikson,E.H.(1968).Identity:Youth and Crisis, Norton, New York.

[5] Erikson,E.H.(1959).Identityand the life cycle.Psychol. lss. 1:1-71.

[6] 人口動態調査“ http://www.mhlw.go.jp/toukei/list/81- 1.htmli ”

[7] A Comparison of Multiclass SVM Methods:Ben Aisen “ http://courses.media.mit.edu/2006fall/mas622j/Projects