2M-02

プロフィール文の属性に基づくツイート分類手法

宮崎 済　　森田 和宏　　泓田 正雄　　青江 順一

徳島大学大学院 先端技術科学教育部

1. はじめに

ソーシャルメディアの普及に伴い，Twitterに代表される SNS を利用した個人の情報発信が活発化している．Twitterはツイートと称される140字以内の短文を投稿できる SNS で，ツイート集合を時系列に並べたものはタイムラインと称される．一人のユーザのツイートが蓄積されたユーザタイムラインはライフログとして機能し，これを解析することによりユーザに関する様々な情報を抽出することができる[1][2].

Twitterには160字のプロフィール文を登録することができる．プロフィール文には一般的に，職業，年齢，趣味嗜好などの属性を端的に表した内容が記載されている．これを読むことで，解析手法を用いずともユーザに関する情報を簡易的に取得することができる．さらに詳しい情報が知りたい場合は，プロフィール文の属性と関連するツイートを直接読めばよい．しかし実際のユーザタイムラインには様々な内容が混在して投稿されているため，可読性が低く，目的のツイートを見つけるには手間がかかるといった問題がある．

ユーザの情報を抽出する関連研究として，馬繰ら[3] はユーザタイムラインから趣味嗜好の推定をおこなう手法を提案している．この手法によって，嗜好を明示していないユーザに対しても推定を可能としたが，精度は低く実用性に問題があると考えられる．また，ユーザを推薦する研究として，原ら[4] はツイート類似度を用いたユーザ推薦システムの手法を提案している．ツイート内容が近いユーザが推薦されるため，目的のツイートを見つけやすくなるがタイムラインの可読性の問題は残っている．そこで，プロフィール文を手掛かりとしてユーザタイムラインを分類することによって，趣味嗜好に関連するツイートのみを読み進められ，目的の情報を得られやすくなると考えられる．

本研究では，プロフィール文に含まれる語の分野属性に基づき，ユーザタイムラインのツイートを分類する手法の提案をおこなう．プロフィール文とユーザタイムラインに対し，分野連想語解析によって属性付けを行い，プロフィール文の属性ごとにツイートを分類，属性別に出力をおこなう．本稿では，提案手法について述べた後に評価実験をおこない，提案手法の有効性の確認と手法改善に役立てる．

A method of tweet classification based on user profile with attributes.
Wataru MIYAZAKI, Kazuhiro MORITA, Masao FUKETA, and Jun-ichi AOE,
Department of Information Science and Intelligent Systems, University of Tokushima

2. 分野連想語解析

分野連想語解析[5] とは，“投手”や“選挙”のように，＜野球＞や＜政治＞といった常識的分野を連想できる単語または複合語(分野連想語と呼ばれる)を使用して文書を解析する手法であり，主に文書分類で利用される．分野連想語解析で得られる分野は階層構造になっているが，本研究では最上位分野を属性ラベルとして用いる．

3. 提案手法

提案手法の概要を図1に示し，以降に手順を述べる．

3.1. プロフィール文解析

プロフィール文に対し，分野連想語解析をおこない属性ラベルを取得する．プロフィール文に対する属性ラベルは，プロフィールラベルとして出力時に用いる．

3.2. ユーザタイムライン解析

ユーザタイムラインのツイートに対し，分野連想語解析をおこなう．3.1 節と同様に属性ラベルを取得した後，プロフィールラベルと照合をおこなう．

3.3. 属性ラベル別ツイート出力

ツイートの属性ラベルとプロフィールラベルの照合結果に基づき出力をおこなう．プロフィールラベルを保有するツイートは属性ラベル別に出力をおこなう．その他のツイートは「ラベル一致なし」，属性ラベルを取得できなかったツイートは「属性なし」として出力をおこなう．

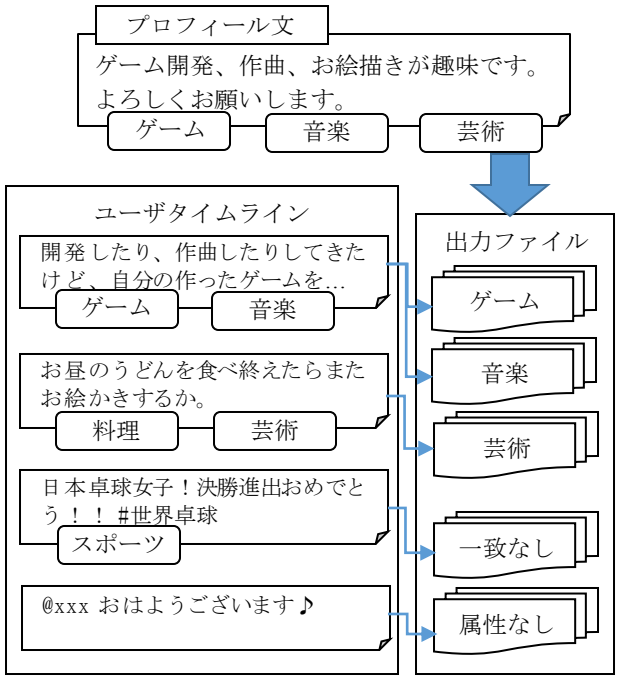

図 1：提案手法の概要

4. 実験

4.1. 実験設定

提案手法の精度評価をおこなった．まず，正解データとして，プロフィール文に「趣味は○○です」のような趣味嗜好を明言する内容を含むユーザ 5 名を選び，それぞれのユーザタイムラインから 100 ツイート，合計 500 ツイートを取得し，プロフィールラベルを含むかどうかの正否を人手で判断したものを用意した．正解データに対し，提案手法による出力結果の精度の計測をおこなった．なお，正否判定には，属性ラベルの種別は考慮しないものとする．

また，取得した属性ラベルの傾向調査をおこなった．上記と同条件のユーザ 10 名を選び，それぞれプロフィール文とユーザタイムラインのツイート約 3,000 件を収集した．提案手法によるツイート分類出力をおこない，各属性ラベルのツイート件数の計測をおこなった．

4.2. 実験結果

精度評価と取得ラベル傾向調査の結果をそれぞれ表 1，2 に示す．表 1 は，ツイート分類精度，再現率，F 値を示している．表 2 は，プロフィールラベルを含むツイート，含まないツイート，属性ラベルを取得できなかったツイートのユーザタイムライン全体に対する出現率を示している．

4.3. 考察

表 1 より，精度，再現率共に良好な結果が得られた．また，表 2 より，ユーザ全体に共通する属性ラベルの傾向は確認されなかった．しかし，一部ユーザを除き，プロフィールラベルを含むツイートの出現率は半数以下となっている．精度の高さと合わせると不要なツイートを大幅に削減できていると言えるため，解読速度向上に有効だと考えられる．また，分類されたツイート内容を調査すると以下のような傾向が確認された．これら 3 項目の分布によりユーザの Twitter の利用傾向を把握することが可能になると考えられる．

- プロフフィールラベルを含む
  傾向： 長文，専門知識，意見・感想
- プロフィールラベルを含まない(一致なし)
  傾向： 長文，自動投稿，意見・感想，日常生活
- 属性なし
  傾向： 短文，挨拶，応答，状況報告

問題点として，属性ラベルと文脈上関連があると考えられる人物名や固有名詞などから，属性ラベルを抽出できていないことが挙げられる．これに対しては，分野連想語解析以外の解析手法の導入や，属性ラベルと名詞の共起頻度などから関連性を推測する必要があると考えられる．また，ユーザタイムライン解析において，「お疲れ様」といった日常的に用いる話し言葉に＜健康＞属性ラベルが付与されているため，偏りが生じている事が確認された．このような例に対しては，分野連想語解析に用いる辞書の拡充や例外ルールを用いることで対応できると考えられる．

表 1：ツイート分類精度(%)

精度	再現率	F 値
94.2	86.4	90.1

表 2：プロフィールラベルの出現率(%)

ユーザ	プロフィールラベル	一致なし	属性なし
A	22.8	35.0	42.2
B	21.3	36.8	41.9
C	15.0	26.1	58.8
D	15.7	31.9	52.3
E	11.8	33.6	54.6
F	50.8	46.4	2.8
G	53.4	22.3	24.3
H	11.6	20.4	68.0
I	11.4	15.5	73.1
J	80.5	17.1	2.3

5. まとめ

本稿では，プロフィール文の分野属性に基づき，ユーザタイムラインのツイートを分類する手法の提案をおこなった．また，精度評価実験をおこない提案手法の有効性の確認をすると共に，分類結果の運用法を考察した．今後は，問題点の改善をおこなうと共に，分類したタイムラインを個別に表示するインターフェイスを開発し，実用性の評価をおこなう．

参考文献

[1] 伊藤淳，西田京介，星出高秀，戸田浩之，内山匡：Twitter と Blog の共通ユーザプロフィール，情報処理学会研究報告，Vol.IFAT-109，No.4，pp.1-8，2013.

[2] 榊剛史，松尾豊：ソーシャルメディアユーザの職業推定手法の提案，日本知能情報ファジィ学会誌 Vol.26，No.4，pp.773-780，2014.

[3] 馬繰美穂，徳久良子，寺嶌立太：ユーザの嗜好と所有物の関係性を用いた属性分析，情報処理学会研究報告，Vol.IFAT-114，No.7，pp.1-6，2014.

[4] 原克彬，中山秦一：呟きに基づいたユーザ推薦システムの提案，第 74 回情報処理学会全国大会，2012.

[5] 辻孝子，泓田正雄，森田和宏，青江純一：複合語の分野連想語の効率的決定法，自然言語処理，Vol.7，No.2，pp.3-26，2000.