深層学習を用いたTwitterからの趣味情報の抽出

Extraction of Interests Information from Twitter Using Deep Learning

若宮 悠希[1*]　砂山 渡[2]　畑中 裕司[2]　小郷原 一智[2]
Yuki WAKAMIYA[1]　Wataru SUNAYAMA[2]　Yuji HATANAKA[2]　Kazunori OGOHARA

[1] 滋賀県立大学大学院工学研究科
[1] Graduate School of Engineering, The University of Shiga Prefecture
[2] 滋賀県立大学工学部
[2] School of Engineering, The University of Shiga Prefecture

Abstract:
In recent years, research to extract personal information from Twitter has been actively conducted. Comments posted on Twitter appear are relatively short and various topics. Therefore comments about a specific interests appear only locally. In this paper, we propose a judgement system whether or not each Twitter user has a designated interest determine after widely learn vocabulary related to specified interests by the deep learning starting from a set of words related to a specific interests.

1　はじめに

近年，インターネットの発展に伴い，Twitter[1]やfacebookといったSNSなどが広く普及することで，オンライン上でのコミュニケーションの場が広く普及してきた．これにより，近しい周囲の人間にとどまらず，距離に縛られない広い範囲の人間と交流を気軽に行うことができるようになった．

これらのサービスを利用するユーザはそれぞれが違う性格や考え方，趣味，嗜好を持つため，交流を行うときは相手の持つ個性が自分のものと合うか，もしくは受け入れられるかどうかが重要となる．そのため，交流相手の個性を十分に理解して受け入れる，もしくは自らが受け入れやすいユーザを選んで交流することが求められる．

交流相手と気が合うかどうかは，同一の話題について興味をもっているかを判断材料とすることができる．Twitterのような不特定多数のユーザが匿名で多様な話題について自由にコメントできるサービスにおいては，検索機能を利用して同じ趣味を持つユーザを選択して交流することもできるが，プロフィールで趣味を明言していないユーザや，サーチワードを含むコメントを投稿したユーザのみしかヒットせず，潜在的に話題に興味を持っているユーザを探すことができないことがあるため，選択の余地が狭まる．

そのため，Twitterに投稿されたコメント集合から，ユーザが特定の趣味を持つか否かを抽出することができれば，交流相手の個性の理解や，新たな交流相手としてユーザを推薦するなど，ユーザ間の交流を手助けが行えることを期待できる．

Twitterから個人情報を抽出する研究は近年盛んに行われるようになってきているが，Twitterにおいては投稿されるコメントが比較的短く，また様々な話題についてのコメントがなされることが多いため，特定の趣味についてのコメントは局所的にしか現れてこない．

そこで本研究では，特定の趣味に関わる単語集合を起点とした深層学習 (Deep Neural Network) により，指定の趣味に関わる関連語彙を潜在的に幅広く学習させた上で，各Twitterユーザが指定の趣味を持つか否かを判定するシステムを提案する．

以下本論文では，2章で関連研究について述べる．3で趣味抽出システムについて述べる．4章で提案システムの有効性の評価について述べ，5章で本論文を締めくくる．

2　関連研究

Twitterに投稿されたコメントから個人情報を抽出する研究は近年盛んに行われてきている．

これまでに，各ユーザの家族構成や所有物，趣味嗜好などを抽出することで個人情報を推定する研究 [1][2]

*連絡先：滋賀県立大学大学院工学研究科 電子システム工学専攻 若宮悠希
〒 522-8533 滋賀県彦根市八坂町 2500
E-mail: of23ywakamiya@ec.usp.ac.jp
[1]https://twitter.com

が行われている．これらの研究では，投稿されたコメント (ツイート) に含まれる単語により個人の情報を抽出する手法を提案しており,「俺のギター」「私の子ども」など，一人称所有格の後に名詞が現れているツイートから，そのユーザの所有物を抽出し,「ギター」であれば「音楽」など所有物が趣味嗜好と関係のあるものならば，所有物を起点にユーザの趣味を抽出する．

また，この他に抽出対象ユーザのプロフィール文やツイートではなく，相互に交流関係のあるユーザのプロフィール文を元に属性を抽出する研究 [3][4] が行われている．この研究では，交流関係のある複数のユーザのプロフィール文に頻繁に出現するものを本人に深く関わりのある単語と仮定して取得することで，これらの単語を元にして対象のユーザに関わりがあると考えられる属性を推定する．

そこで本研究では，自身の学士の研究 [5] を元に，抽出したい趣味を利用者があらかじめ決定して関連単語を与えることにより，趣味と関連する単語として幅広い語彙を網羅した学習を行うことで，推定対象のユーザのツイート集合のみから，指定の趣味を持つか否かの判断を行う．

3 深層学習を用いた趣味抽出システム

3.1 システムの構成

本研究で提案する深層学習を用いた趣味抽出システムの構成を図 1 に示す．

まず，あらかじめ深層学習を用いて構築した各趣味の趣味抽出ネットワークを元に，推定対象の Twitter ユーザのツイート集合に各趣味が現れているかを抽出する．抽出結果を可視化インタフェースを利用して提示することで，システム利用者が結果を解釈する支援とする．

3.2 深層学習による趣味抽出ネットワークの学習

本研究では，推定対象ユーザが指定の趣味に興味を持っているかを判断するために，深層学習による趣味抽出ネットワークを構築，これを元にツイート集合を分類する．

深層学習とは，機械学習の一種で入力と出力の関係をパターンとして学習したネットワークを構築する手法である．構築されたネットワークにより，人間に与えられない細かな分類規則によって新たなデータを分類することができる．

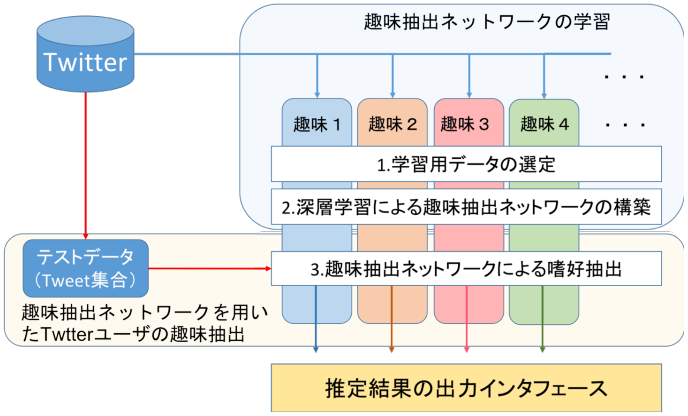

図 1: 趣味抽出システムの構成図

本研究では，深層学習ライブラリの 1 つである Deep Learning for Java(DL4J)[2]を利用して， Deep Neural Network を扱う．

3.2.1 深層学習に利用するデータ集合

本研究では，趣味抽出の対象を Twitter ユーザに定め，各ユーザの投稿したツイート集合を深層学習によって分析し，指定した趣味が文章中に現れているかを抽出する．Twitter ユーザの趣味抽出を行う際は，対象ユーザの投稿した複数のツイートを 1 件ずつ推定していくことで，全体のうち何%に趣味が現れているかを元に最終的な判断をする．

3.2.2 深層学習に利用するデータの収集

深層学習に用いる入力データは，推定対象と同じく，Twitter から収集したツイートを利用する．抽出対象とする趣味が文章に現れているツイートを正例，対象とする趣味が文章に現れていないツイートを負例として正解ラベルをつける．これらのツイート集合を入力データとして，深層学習により趣味抽出ネットワークを構築する．

ここで，ツイートに指定の趣味が現れているか否かは，指定の趣味に関連のある単語集合を設定し，いずれかの単語が文章中に出現しているか否かを判断する基準とする．例えば,「野球」に関していえば「阪神タイガース (プロ野球チーム名)」や「ホームラン (野球用語)」などが考えられる．この単語集合を本研究では「初期単語」と定義する．

Twitter REST API[3]を用いてツイートを収集し，指定の趣味ごとに設定した初期単語を含むツイート，含まないツイートを選別して正例，負例を選択する．

[2]https://deeplearning4j.org
[3]https://developer.twitter.com/en/docs

深層学習の入力データや推定対象の文章は，それぞれ文章から BoW(Bag of Words) に変換して利用する．BoW とは全文章中に出現する単語を並べ，各文章での単語の出現頻度をベクトルで表現したものである．また，抽出対象の趣味特有の表現を学習したいため，使用する単語は 15 ツイート以上に出現した名詞のみとし，その中でも判定に関わらないと考えられる単語はあらかじめ除去する．除去する単語としては,「リツイート」「リプライ」など，Twitter で趣味関係なく広く使われる言葉や，初期単語の有無に関わらずツイートに出現する単語となる．初期単語を含む正例データと，初期単語を含まない約 220 万のツイート集合との間で，出現した単語全てについてカイ 2 乗検定を行い，有意水準 5 %を下回る単語以外を全て除去する．

3.2.3 趣味抽出ネットワークの構築

指定した各趣味に対し，正例，負例の入力データを元に深層学習を行い趣味抽出ネットワークを構築する．入力として BoW を用いることから，入力層ノード数は学習用データ中に出現する名詞の総数となるため，抽出対象となる趣味ごとに入力層ノード数が異なる．その他の深層学習のパラメータは，全て以下のもので統一する．中間層数:3，中間層ノード数:100，中間層活性化関数:Relu，出力層ノード:数 2，出力層活性化関数:Softmax，エポック数:50，ミニバッチサイズ:256，L1 正則化 (係数 0.01)，Adam 利用．

今回，深層学習を行う上で，趣味抽出の結果が後述の評価において最も高い精度となったパラメータを採用した．

3.3 趣味抽出ネットワークを用いた Twitter ユーザの趣味判定

構築した趣味推定ネットワークを利用して対象ユーザのツイートから趣味抽出を行う．趣味抽出を行う際には，推定対象の Twitter ユーザが投稿した複数のツイートの集合を対象とし，ツイート 1 件を 1 データとして，文章中に趣味が現れているかどうかを 2 値で判定する．次に，趣味が強く現れていると正判定されたツイートが，全体のツイート集合に対しどの程度の割合で現れたかを算出してグラフ出力する．抽出結果を提示するための可視化インターフェースとして，統合開発環境である TETDM[6] を利用する．

この時，最終的にユーザが指定の趣味に興味を持っているかどうかは，あらかじめ定めた閾値を超えているかどうかで判断する．指定する趣味の範囲や，どれだけ一般的かにより，文章からの抽出されやすさが異なるため，閾値を一般的な Twitter ユーザの判定結果

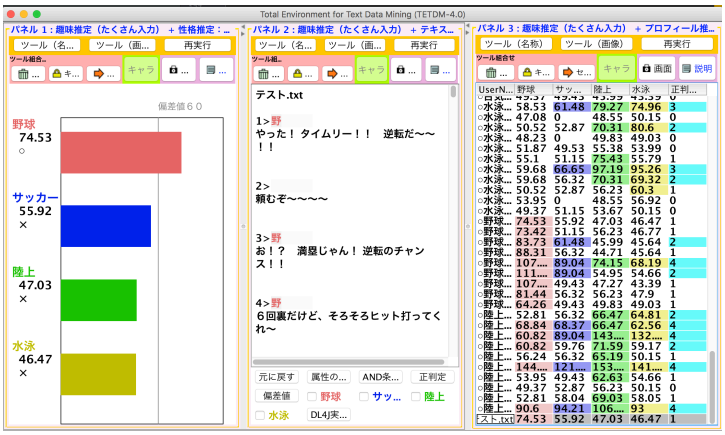

図 2: 趣味抽出システムの使用例

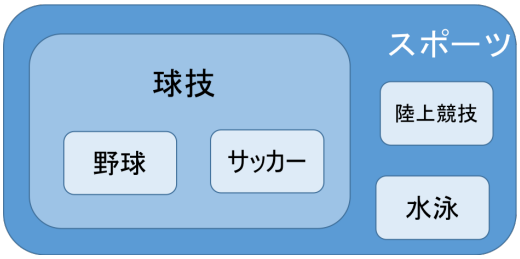

図 3: 趣味の範囲

よりも正判定ツイートの割合が少し高くなるように設定する．ランダムに収集した 15435 ユーザに対し上記の趣味判定を行い，各ユーザの判定結果から正判定ツイート割合の偏差値を算出し，偏差値 60 となる割合を閾値とする．

3.4 趣味抽出システムの使用例

趣味抽出システムの使用例として，Twitter 上での交流相手として，趣味の合うユーザを探し出したい場合，提案システムを利用して，自らの指定した複数の趣味を持つユーザを探し出す例を挙げる．

TETDM を利用した趣味抽出結果の提示例を図 2 に示す．取得した複数ユーザのツイート集合各 100 件程を同時に入力し，趣味抽出システムを起動する．図 2 に示すように，TETDM 上に各ユーザが指定の趣味をどれだけ持つかを正判定ツイート割合の偏差値により表示され，選択したユーザの抽出結果がグラフがツイート集合とともに表示される．図 2 の例では，後述する「野球」「サッカー」「陸上競技」「水泳」の趣味が現れているユーザを，複数入力したユーザの中から抽出しており,「スポーツ」に興味を持つユーザを選択する支援を行える．

3.5 指定の趣味の決め方と抽出方法

本研究では，指定の趣味の関連語彙を学習させることで，対象ユーザのツイート集合に含まれる単語より趣味に興味があるか否かを抽出する．しかし，推定対象となり得る趣味は図 3 に示すようにそれぞれで範囲が異なる．趣味の範囲が広くなる程に内包する属性が多くなるため，初期単語の決定や関連語彙の十分な学習が困難となる．一方で，狭い範囲の趣味を学習する場合，関連語彙を学習しすぎて，少し広い範囲まで抽出してしまう場合がある．例えば,「野球」についての学習を行った結果,「優勝」「攻撃」「応援」「監督」などの関連語彙を網羅する学習を行うが，これらの単語は「スポーツ」全般で利用される言葉である．

そのため本研究では，抽出対象よりも少し狭い複数の趣味を起点として提案手法を用いることで，想定する範囲の趣味を抽出する．

4 趣味抽出システムの評価

3 章で述べた提案手法を評価するため，実際に趣味抽出を行い，抽出可能な趣味の種類や，判定方法などの評価結果を示す．

4.1 提案システムの評価

提案手法を評価するため，一般的な趣味の 1 つである,「スポーツ」趣味を「野球」「サッカー」「陸上競技」「水泳」の各趣味を起点として抽出する．各趣味についてそれぞれ初期単語を設定し，1 つずつ趣味抽出ネットワークを構築し，これらからテストユーザから正しく「スポーツ」趣味を抽出できるかを評価する．

各趣味について設定した初期単語を表 1，それを起点に収集した学習用データ数，入力ノード数 (利用する単語数) を表 2 に示す．また,3.3 章で述べた，各趣味における最終的な正判定の閾値を表 3 に示す．

「スポーツ」評価用として，正解ラベルが正，負のものを合わせて 168 のテストユーザを収集した．収集基準，内訳は以下の通りであり，ユーザの投稿ツイートを読み，受けた印象を元に人手で収集した．

負 特にスポーツに興味がなさそうなユーザ：100

正 各スポーツに興味がありそうなユーザ：68
「野球」9,「サッカー」10,「陸上」10、
「水泳」11,「ゴルフ」10,「卓球」10、
「テニス」5,「ラグビー」2,「合気道」1、
「スケート」1

これらのテストユーザを用いて，各趣味を起点にした場合における「スポーツ」判定を評価した．評価結果として，正判定についての Precision，Recall，F 値を表 4 に示す．

表 4 の結果では，どの趣味を起点として「スポーツ」判定を行っても，高い Precision と低い Recall が目立つことから，提案手法では，誤判定が少ないものの正判定の基準が厳しく設定されていることがわかる．

また，負判定された「正」ユーザの中には，正判定ツイート割合が閾値にわずかに届かなかったものがあった.「入力ツイートの内，正判定ツイート割合が決定した閾値を超えている」ことが最終的な正判定の条件となるため,「スポーツ」に確かに興味があるユーザについても，閾値を下回れば負判定されてしまう．このことから，閾値の設定もさらに慎重に行うべきであることがわかる．

1 つの趣味を起点とした学習により，起点趣味のみに限らず周辺の趣味までの関連語彙を学習し,「スポーツ」の理想的な範囲を網羅できているかを，表 4 の内訳として表 5，6，7，8 により確認した．

「サッカー」起点なら「サッカー」,「野球」起点なら「野球」に興味を持つユーザといったように，起点趣味に興味のあるユーザは高い Recall で判定され，起点趣味に興味を持たないその他のユーザのいくつかを正判定している．ここから，1 つの起点趣味からの学習により，少し広い範囲の関連語彙までを網羅できることがわかる．

一方で，各起点趣味により性判定されやすい趣味が異なる．これは，起点趣味により初期単語が異なることから，学習した語彙が異なることや,「スポーツ」全体に関わる単語を学習できていても，他趣味の固有名詞まで網羅できないことが原因だと考えられる．このことから，起点趣味 1 つのみから「スポーツ」を網羅するのではなく，複数の起点趣味を組み合わせることによって「スポーツ」判定を実現することを考える．

4 つの起点趣味からの学習により構築した趣味抽出ネットワークそれぞれにより趣味抽出を行い，指定した数以上の回数で正判定をされたユーザを，最終的な「スポーツ」趣味を持つユーザであると判断することで，それぞれの起点趣味による学習結果を補い合う．テストユーザを用いた以上の手法の評価結果を，表 9 に示す．

正判定回数を高く設定した場合には，多くの起点趣味で共通して学習した語彙を多く使うユーザが正判定されるなど，判定が更に厳しくなるが，低く設定した場合，それぞれが学習した異なる語彙が判定結果を補い合うことでより適切に判定を行えるようになった．

一方で，どの起点趣味による趣味抽出でも，一度も正判定がつかなかった「正」ラベルのテストユーザが 17 あり，これらのユーザはツイートから受ける印象として，確実に「スポーツ」趣味を持っているのにも関

表 1: 趣味ごとの初期単語

趣味名	初期単語
野球	「ホームラン」「ピッチャー」「バッター」
サッカー	「ゴールキーパー」「フォワード」「ミッドフィールダー」
陸上競技	「ベリーロール」「クラウチング」「背面跳び」「はさみ跳び」「長距離走」「短距離走」
水泳	「クロール」「平泳ぎ」「背泳ぎ」

表 2: 趣味ごとの学習用データ数と入力層ノード数

趣味名	学習用データ数	入力層ノード数
野球	20004(正例:10002)	960
サッカー	20004(正例:10002)	1186
陸上競技	20004(正例:10002)	865
水泳	20004(正例:10002)	1002

表 3: 趣味ごとの閾値 (偏差値 60 となる正判定ツイート割合)

趣味名	閾値
野球	0.131
サッカー	0.081
陸上競技	0.148
水泳	0.185

表 4: 各趣味を起点にした場合における「スポーツ」判定評価結果

趣味名	Precision	Recall	F 値
野球	0.941	0.471	0.627
サッカー	0.935	0.426	0.586
陸上競技	0.911	0.603	0.726
水泳	0.842	0.471	0.604

表 5: 「野球」起点での趣味判定結果の内訳

	「正」予測数	正解ユーザ数	Recall
野球	9	9	1.00
サッカー	6	9	0.67
陸上	5	10	0.50
水泳	0	11	0.00
ゴルフ	4	10	0.40
卓球	5	10	0.50
テニス	1	5	0.20
その他	2	4	0.50

表 6: 「サッカー」起点での趣味判定結果の内訳

	「正」予測数	正解ユーザ数	Recall
野球	3	9	0.33
サッカー	6	9	0.67
陸上	5	10	0.50
水泳	2	11	0.18
ゴルフ	2	10	0.20
卓球	6	10	0.60
テニス	3	5	0.60
その他	2	4	0.50

表 7: 「陸上競技」起点での趣味判定結果の内訳

	「正」予測数	正解ユーザ数	Recall
野球	1	9	0.11
サッカー	4	9	0.44
陸上	9	10	0.90
水泳	5	11	0.45
ゴルフ	7	10	0.70
卓球	8	10	0.80
テニス	3	5	0.60
その他	2	4	0.50

表 8: 「水泳」起点での趣味判定結果の内訳

	「正」予測数	正解ユーザ数	Recall
野球	1	9	0.11
サッカー	4	9	0.44
陸上	5	10	0.50
水泳	4	11	0.36
ゴルフ	7	10	0.70
卓球	7	10	0.70
テニス	1	5	0.20
その他	2	4	0.50

表 9: 複数の起点趣味を組み合わせた「スポーツ」判定評価結果

正判定回数	Precision	Recall	F 値
4	1.000	0.265	0.419
3	1.000	0.368	0.538
2	0.907	0.574	0.703
1	0.839	0.765	0.800

表 10: 判定閾値 52 の場合の複数の起点趣味を組み合わせた「スポーツ」判定評価結果

正判定回数	Precision	Recall	F 値
4	0.975	0.574	0.722
3	0.842	0.706	0.768
2	0.803	0.838	0.820
1	0.626	0.912	0.743

わらず，正判定ツイート割合が閾値に到達していないため，誤判定を受ける結果となっていた，

前述の通り閾値の適切な決定方法を調査するために，正判定ツイート割合が偏差値 40 から 60 まで閾値を変化させて，表 9 の評価を再度行った．この時，最も良い結果を得られた偏差値 52 の結果を表 10 に示す．

閾値を少し下げて適切な値に調整することにより，わずかに閾値に届いていなかった「正」ラベルテストユーザを正しく判定することができた．また，閾値を甘くすることにより,「スポーツ」趣味のないユーザを誤判定してしまうリスクも高まったが，複数の起点趣味で正判定されることを条件に加えることで，Precision を大きく下げることなく，Recall を高くすることができ，より適切な判定結果を得ることができた．これらの結果から，提案手法を用いる際には複数の起点趣味を用意し，それらを組み合わせて趣味抽出を行うべきであると考えられる．

しかし，閾値や組み合わせる起点趣味の数の設定方法などは未だ検討途中であるため,「スポーツ」以外の判定も合わせて行うことで明確なルールを決定する必要がある．また，今回の評価に用いたテストユーザは，個人の主観により収集したものとなるため，複数人での選定等を行い，より信頼性の高いデータを元に評価実験を行う予定である．

5 おわりに

Twitter ユーザを対象として指定の趣味抽出を行い，抽出結果を利用者に提示するシステムの作成を目的として研究を行っている．初期単語と定義した，各趣味に関連の深い語彙を設定することで，これを起点に深層学習により周囲の関連語彙を潜在的に学習することで，対象ユーザの投稿ツイート集合から趣味を抽出して提示するシステムを開発した．

起点となる趣味を元に関連語彙を学習することにより，起点よりも広い範囲の補完を試みるため，一般的な趣味の 1 つである「スポーツ」についての評価を行った．その結果，複数の趣味を起点として組み合わせることにより，学習結果を補い合いより適切な判定結果を得ることができた．

今後の課題として，学習の際に設定する初期単語や起点趣味などの適切決定方法の考察や，テストユーザの増加などにより，更に踏み込んだ評価を行うことで，提案手法の有効性を高めていくことを目標としていきたい．

参考文献

[1] 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研究報告情報基礎とアクセス技術 2014-IFAT-114,pp.1-6 (2014)

[2] 那須川哲也, 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,pp.952-955 (2013)

[3] 上里 和也, 浅井 洋樹, 山名 早人:Personalized PageRank を利用した網羅的 Twitter ユーザ属性推定, DEIM 2016 第 8 回データ工学と情報マネジメントに関するフォーラム D2-2 (2016)

[4] 鈴木 祥平, 池田 拓生, 倉田 陽平, 石川 博:Twitter のユーザプロフィールを用いた観光地の類型化, DEIM 2016 第 8 回データ工学と情報マネジメントに関するフォーラム A2-1 (2016)

[5] 若宮 悠希, 砂山 渡, 畑中 裕司, 小郷原 一智:深層学習を用いた Twitter ユーザの性格推定 2018 年度人工知能学会全国大会 (第 32 回) 3F1-OS-12a-03 (2018)

[6] 砂山渡, 高間康史, 徳永秀和, 串間宗夫, 西村和則, 松下光範, 北村侑也: 統合環境 TETDM を 用いた社会実践, 人工知能学会論文誌 32 巻 1 号,pp.NFC-$A_1 - 12$(2017)