Twitter からのユーザ情報抽出およびプロフィール推定

的埜, 孝宏 / MATONO, takahiro

(出版者 / Publisher)
法政大学大学院情報科学研究科
(雑誌名 / Journal or Publication Title)
法政大学大学院紀要. 情報科学研究科編 / 法政大学大学院紀要. 情報科学研究科編
(巻 / Volume)
9
(開始ページ / Start Page)
149
(終了ページ / End Page)
154
(発行年 / Year)
2014-03
(URL)
https://doi.org/10.15002/00010537

Twitterからのユーザ情報抽出及びプロフィール推定

User Information Extraction and Profile Estimation from Twitter

的埜　孝宏

法政大学大学院情報科学研究科情報科学専攻

E-mail: takahiro.matono.5w@hosei.ac.jp

***Abstract*—In these days many people use Social Network Services (SNS) such as Facebook, Line and Twitter though computers and smartphones. By mining the SNS data, a lot of applications can be made, for instances, special event finding, people's social relations, market planning, recommendation, etc. One of important researches is to automatically or semi-automatically build a model for a user, based on analysis of user's SNS data. Generally, user modeling is to estimate the hobbies, preference, behavior and state from the user's various data. This research is aimed at estimations of a user's profile by using morphological analysis to extract the user's information from Twitter data. The profile in our study consists of three parts about a user's address, career and interests. Our research has been focused on approaches to improve accuracy in estimations of the user's address, career and interests. To improve the accuracy of address estimation, both address dictionary and associative words are used to count address occurrence frequency. The career estimation is based on extraction of typical words associated with a career e.g., a working adult, a student or a housewife. The interest estimation is more difficult due to the diversity of possible interests for different people. Our interest estimation has been made with three approaches, Jaccard coefficient and interest book. About 30,000 tweets are used for data analyses and user profiling as well as tests and evaluations of the proposed approaches for address, career and interest estimations.**

Keywords—Twitter; user profile; address; career; interest; morphological analysis; estimation; accuracy

I.　序論

近年，スマートフォンやコンピュータといった情報通信機器の普及やネットワーク環境の拡大にともない様々な人たちのライフログを抽出することが可能になった．抽出されたライフログからユーザの行動状態や趣味・嗜好などを推定する研究が多くおこなわれている．例えば，小林らの研究[1]では，携帯電話に搭載されている加速度センサの値からユーザの“歩行，走行”などの行動状態を推定している．

一方で，現在スマートフォンの普及にともない Facebook, LINE, Twitterといった SNS(Social Network Service)を利用するユーザが増加している．日本の SNS 利用者は 2015 年末には 6,321 万人の利用者が見込まれている．中でも Twitter は，1 分間に 625 アカウント増加しており，2012 年には 5 億アカウントを突破するといった爆発的な増加をしている．また，Twitter は 1 日にツイートされる数は 5 億ツイート，毎日利用するユーザ数は 1 億人，1 月の間に利用するユーザ数は 2 億 1830 万人と使用頻度も高い[2].

加えて，Twitter にはユーザが匿名で使用することができることから，生の声を聞くことができ，肯定的な意見よりも否定的な意見を参考にすることができる[3]．これらのことから，Twitter ユーザの意見を取り入れ，マーケティング分野で活用する研究がおこなわれている．マーケティング分野で使用するためには Twitter ユーザの意見だけではなく，ユーザの性別，年齢，職業などのユーザ属性が必要である．そこで，先行研究では，これらのユーザ属性の推定がおこなわれている．

本研究ではこれらの Twitter の利点から，Twitter を基にしてプロフィールを推定する．また，本研究では先行研究のマーケティング分野で利用することや共通点が多いユーザを推薦することで人々が繋がることを想定したプロフィールの推定をおこなう．具体的には，ユーザ属性のうち「居住地や出身地」「職業」が同じであれば共通の話題が多く産まれる可能性が高いことが考えられる．また，ユーザの「興味」を持っている内容が同じであれば共感できる可能性が高いと考えたため，本研究では「居住地や出身地」「職業」「興味」の 3 つの項目を記載することにした．従来の研究では，これらのユーザ属性に対して，十分な精度が得られていないことから，本研究では，推定精度の向上を目的とする．また，本研究では，これらの 3 つの項目の推定によるプロフィールの自動生成を目指す．

本論文の構成は以下の通りである．第Ⅱ章では，本研究に関連する研究の紹介をする．第Ⅲ章では，本研究の大まかな流れと研究全体で利用した技術について述べる．第Ⅳ，Ⅴ，Ⅵ章では，プロフィールに記載する「関連地」「職業」「興味」についてそれぞれの推定手法及び実験内容・結果など詳しく述べる．第Ⅶ章では，本研究により作成されたプロフィール例及び作成されたプロフィールについて実施したアンケート調査の結果について述べる．第Ⅷ章では，本研究のまとめについて述べる．

II.　関連研究

この章ではまず本研究のプロフィールに記載する内容について詳しく述べる．そして，関連する手法を述べた後に本手法の特徴について述べる．以下に 3 つのセクションに分けて述べる．

A. ユーザプロフィール

ログを用いてユーザのプロフィールを推定及び作成する研究は多くおこなわれている．例えば，Susan Gauch らの研究[4]では，ユーザのパソコン上でのログ(ブラウザーキャッシュや検索履歴など)を使用してユーザのプロフィールを作り出す研究がある．池田らの研究[5]では，Twitter のログを用いて，ユー

ザ属性の推定をおこなっている．池田らは，統計的に偏って現れる特徴的な語(十代であれば“バイト”，男であれば“俺”など)を用いて SVM で推定をおこなっている．平野らの研究[6]も同じく，Twitter のログを用いてユーザ属性の推定がおこなわれている．平野らは，ユーザ属性間における関連性に注目をし，推定を行っている．以下の図 1 に池田ら(左)と平野ら(右)によって推定されたユーザ属性をそれぞれ示す．

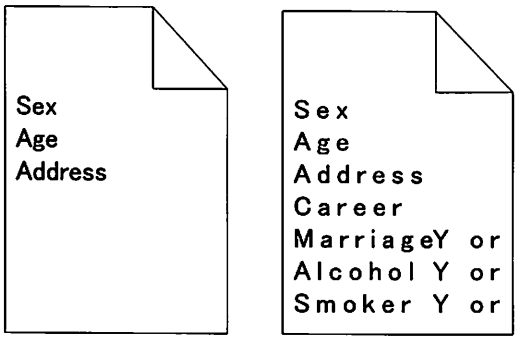

図 1. 池田らの項目(左)，平野らの項目(右)

上記の図 1 のユーザ属性は，Twitter ユーザの意見を取り入れ，マーケット活動に活かす目的からユーザ属性の推定がおこなわれている．そのため，マーケティングのアンケート調査を参考に，ユーザの住んでいるところを関東圏や近畿圏など「居住域」として広く推定している．また，「職業」においては“会社員，自営業/個人事業，公務員，団体職員，アルバイト/パート，専業主婦・主夫，大学生，高校生，無職，その他”の 10 種類に分類をおこなっている．

本研究は，人々が繋がることを想定した時，「居住域」という広い区域では共通の話題が出にくいと考えたため，もっと局所的な推定が必要であると考えた．そのため，都道府県単位で推定をおこなうことにした．また，人々が繋がるためには「居住地」だけではなく「出身地」の情報も有用であると考えたためユーザに関連深い土地として，2 つの場所をまとめて「関連地」として記載することにした．また，上記の「職業」を大まかにまとめると“社会人，学生，主婦”の 3 種類であり，本研究では各ユーザ属性の精度を向上させることが目的のため，ユーザの「職業」を“社会人，学生，主婦”の 3 種類に分類することで精度向上を試みた．加えて，本研究ではユーザ属性以外にもユーザの「興味」を推定することが必要であると考えた．そのため，本研究では幅広いユーザの「興味」の中から，あらかじめ男女共通の「興味」として“食べ物”，男性向けの「興味」として“スポーツ”，女性向けの「興味」として“ファッション”の 3 種類を用意した．

B. 様々な推定手法

榊らの研究[7]では，川原らの「格助詞＋動詞」の組み合わせで，格助詞に係る名詞の意味が限定されるという研究[8]を参考にして，特定の格助詞「で」と特定の動詞「会う」を用意することで，格助詞に係る名詞の意味を「場所」または「理由」の 2 つに限定をおこなっている．呉らの研究[9]では，あらかじめ分野ごとに分類(政治，社会，経済，国際，スポーツ)された文章群から名詞を抽出し，分野内及び分野間の名詞の出現頻度をもとに分野関連語辞書を作成し，後に新聞記事など一般的な文章を自動的に分類する手法が紹介されている．小川らの研究[10]では，名詞の出現頻度から「特徴的な情報」を導き出すため，一定期間に出現頻度の激しく変化する語に着目し，その語を「特徴的な語」として抽出をおこなう手法が提案されている．また，「ソーシャルブックマークを基にした Twitter ユーザの興味語抽出・推薦手法の提案と評価」[11]という論文では，ソーシャルブックマークからユーザの興味語を抽出する手法について述べられている．この論文では，単語の共起(共に起こる事象)関係からユーザの「興味」を推定している．

C. 本手法の特徴

本手法の特徴は，ツイートの文中には位置情報が付加された文字列が大量に得られることが示されていることから[12]，「関連地」を推定する場合，全国の地名名詞(都道府県名，市名，東京 23 区名)とユーザのツイート文中に含まれる位置情報とのマッチングおこない，地名名詞の出現頻度を調べることでユーザの「関連地」の推定をおこなっている点である．さらに，本手法では，ツイート文中に含まれる位置情報が他の意味(人名の香川や石川など)の語と誤ってマッチングしないように榊らと川原らの研究を取り入れ改善した点にある．また，ユーザの「職業」を推定する場合，呉らの手法は，新聞のジャンルにおける特徴語から自動的に文章を分類しているが，本研究では各職業において特徴的な語(社会人であれば“仕事”など)から，ユーザの「職業」を推定している．さらに，全ツイートから推定をおこなうと，ユーザの過去の「職業」が推定されることから，一定の期間を定めて抽出し，過去の「職業」が推定されないようにした点である．また，本研究ではユーザの「興味」を推定するために，あらかじめユーザが「興味」を示しそうなジャンルを用意した点が特徴的である．また，小川らがおこなった手法を Twitter において応用しユーザの「興味」を抽出することを目指した点がある．これらの推定手法については第Ⅳ，Ⅴ，Ⅵ章で詳しく述べる．

III. 研究概要

この章では本研究の流れ及び研究全体で利用した技術である形態解析技術，形態素解析するために利用したテキストマイニングツール KH Coder について，2 つのセクションに分けて述べる．

A. 本研究の流れ

プロフィールに記載する事項のうちユーザの「関連地」を推定するために，政府統計の総合窓口というサイトの都道府県・市町村別統計表[13]を参考に全国の地名名詞を用意した．また，ユーザの「職業」を推定するために，あらかじめ“社会人，主婦，学生”から 10000 ツイートずつ抽出・形態素解析をおこない，各職業において特徴的な語を抽出し用意した．そして，ユーザの「興味」を推定するために必要な各ジャンルに関連深い名詞を 30000 ツイート(各職業のツイート)から形態素解析をおこない抽出し用意した．そして，これらを用いて，推定するユーザのツイート(1 人のユーザから抽出可能なツイート数は 3200 ツイート)とマッチングをおこない出現頻度の高い語を調べ，各項目において重要な語として抽出する．尚，ユーザの「興味」を推定するために用意したデータは，他のユーザの「興味」を推定する手法にも用いた．以下の図 2 に各項目で用

意したデータとマッチングをおこない，出現頻度の高い語を調べ，プロフィールが作成されるまでの一連の流れを示す．

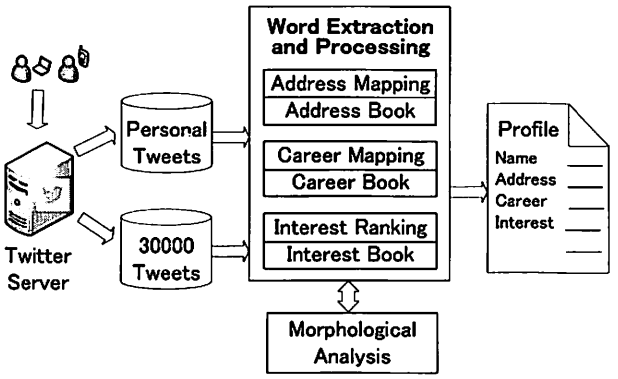

図 2. プロフィール推定の流れ

B. 形態素解析技術

形態素解析とは文章中の語句を言語で意味を持つ最小単位である形態素ごとに分解を行い，「MeCab」，「茶筅」といった品詞体系(名詞，形容詞，動詞など)ごとに語句を分けて登録してある辞書とのマッチングにより品詞分解をすることである．本研究では，形態素解析をおこなうため樋口氏により開発されたデータマイニングツール「KH Coder」を使用した．KH Coder は形態素解析をした単語を品詞ごとに分類し，Excel 形式で各単語の出現頻度を出力する機能がある．また，“食べる”や“食べた”などの同じ意味であるが，違う形式の単語をまとめてカウントする機能もある．本研究では，「職業」「興味」を推定するために必要な特徴語を抽出するために利用した．尚，本研究では，KH Coder に初期に搭載されている品詞体系辞書「茶筅」を使用した．以下の図 3 に抽出したツイートを形態素解析にかけ Excel 形式で出力した結果を示す．

	A	B	C	D	E
1	名詞		サ変名詞		形容動詞
2	バイト	180	一緒	123	大丈夫
3	自分	110	お願い	99	マジ
4	久しぶり	76	話	67	好き
5	学校	74	テスト	65	無理
6	生年月日	69	勉強	64	普通
7	電車	67	試合	60	暇
8	感じ	64	授業	52	残念
9	楽しみ	62	誕生	52	いや
10	大学	56	予定	48	ダメ
11	最後	48	期待	42	大好き
12	疲れ	47	合宿	36	大変

図 3. KH Coder による形態素解析結果

IV. 関連地推定

この章ではまず，ユーザの「関連地」を推定するために用いた手法について詳しく述べる．そして提案する手法による実験と結果について述べる．

A. 関連地推定手法

ユーザの「関連地」を推定するために，まずツイート文に含まれる位置情報と地名名詞(都道府県名 47 個，全国の市名 789 個，東京の区名 23 個を用いた)をマッチングすることで，それぞれ出現回数をカウントする．そして，各都道府県名の出現回数とそこに含まれる市名・東京 23 区名の出現回数の合計を求め，出現頻度の最も高い都道府県名をユーザの「関連地」として推定する．尚，本研究では，ユーザがツイートする時，位置情報に「都」「府」「県」「市」「区」を付けて発言する可能性が低いことが考えられたため，地名名詞からこれらの語を取り除いたものとマッチングすることで抽出率を上げることにした．ただし，全部ひらがなである地名名詞(さぬき市やあま市など)と漢字 1 文字である地名名詞(北区や光市など)は，他の意味の語と誤ってマッチングする可能性が高いと予測されたため，本研究ではこれらの言葉からは「市」「区」を取り除かないことにした．加えて，都道府県名と市名・区名が一致する地名名詞は，どちらか一方をカウントするようにした．本研究では，この手法を「関連地マッチング手法」と述べる．

しかし，単にマッチングを行った「関連地マッチング手法」の場合上記で述べたように，他の意味の語と誤ってマッチングする可能性がある．“News ポストセブン”[14]の記事によると 47 都道府県の名所のうち名字として存在しないのは“北海道，愛媛，京都，沖縄”の 4 つのみである．そこで本研究では，榊らと川原らの研究を参考に「関連地マッチング手法」で使用された地名名詞に格助詞「で」「にて」「から」や Twitter における特徴的な語「なう」などの特定の語を付加する．本研究ではこの手法を「関連地パターンマッチング手法」と述べる．この手法においては，逆に地名名詞に何も付加していない語は抽出できないという問題点が存在するため，上記で取り除いた語や地名名詞のあとに続きやすい語も合わせて付加することで抽出率を上げることにした．この手法は，「関連地マッチング手法」に比べてマッチングする個数が減少するというデメリットが考えられる．

そこで，本研究では，これら 2 つの手法において推定される結果から良い手法を用いて，ユーザの「関連地」を推定することにした．以下に，「関連地パターンマッチング手法」で地名名詞に付加した語を表 1 に示す．

表 1. 地名名詞に付加した語

格助詞	「で」「にて」「から」
Twitter における特徴語	「なう」
その他	「都」「府」「県」 「市」「区」「駅」 「帰」「行」「着」

B. 関連地推定実験・結果

本研究では，推定された「関連地」がユーザの「出身地」か「居住地」であれば正しく推定されたものとするため，あらかじめユーザの「出身地」と「居住地」を調査した．そして，抽出したユーザのツイート数に応じて推定精度に違いが出ることが考えられたため，0~1000 ツイートのユーザを 6 人，1000~2000 ツイートのユーザを 7 人，2001~3200 ツイートのユーザを 7 人それぞれ実験対象にした．また，“東京，大阪”などの大都市は他の地域に比べて出現しやすいのではないかと考えた．そのため，実験対象にした 20 人のユーザの中には“東京，大阪”などの大都市が含まれていないユーザが 11 人含まれている．以下に「関連地マッチング手法」，「関連地パターンマッチング手法」の推定精度を表 2 と表 3 にそれぞれ示す．

表 2.「関連地マッチング手法」推定結果

	正	誤
ツイート数	関連地	関連地以外
0~1000	6	0
1001~2000	6	1
2001~3200	6	1
合計	18(90%)	2

表 3.「関連地パターンマッチング手法」推定結果

	正	誤
ツイート数	関連地	関連地以外
0~1000	5	1
1001~2000	7	0
2001~3200	7	0
合計	19(95%)	1

これらの結果からどちらの手法を用いても高い精度で推定することが可能なことが確認された．しかし，推定を誤った理由はそれぞれ異なり，「関連地マッチング手法」では「関連地」である「居住地」と「出身地」以外の場所が推定されたのに対して，「関連地パターンマッチング手法」ではユーザのツイート数が少なかったことから抽出可能な地名名詞の個数が少なかったことが原因で誤った場所を推定したと推測される．また，あるユーザで実験をおこなったところ，「関連地マッチング手法」では“香川”という単語が 130 回出現していたが「関連地パターンマッチング手法」では 1 回であった．そこで“香川”を調べたところ地名の“香川”ではなくほぼ全部が人名の“香川”であることが分かった．そこで，本研究では推定精度と他の意味の語と誤ってマッチングする可能性が極めて低い「関連地パターンマッチング手法」を用いて「関連地」を推定することにした．

V. 職業推定

この章ではまず，ユーザの「職業」を推定するために用いた手法及び推定に必要な各職業における特徴語の抽出方法について述べる．そして，実験方法と推定精度を述べた後に．さらに精度を向上させる手法を提案し，実験及び推定精度についてのべる．

A. 職業推定手法及び特徴語抽出

ユーザの「職業」を分析するために本研究では事前に各職業から 10000 ツイートずつ抽出・形態素解析を行い，各職業における複数の特徴的な語を抽出する．そして，ユーザの「職業」を推定する際に，各職業における複数の特徴的な語の出現回数をそれぞれカウントし，職業ごとに複数の特徴的な語を合計する．そして，出現頻度の最も多い職業をユーザの「職業」として推定する．本研究では，各職業における特徴的な語を各職業との関連性の高さと出現頻度の高さを考慮して選択した．各職業における特徴的な語を調べたところ“社会人”に関連性が高く，出現頻度が 3 番目の語と 4 番目の語に大きな出現頻度の違いがあったため，各職業から 3 語ずつ抽出をおこなった．そして，これらの特徴的な語から 2 語を用いた場合と 3 語を用いた場合を事前に比較実験をおこなった．結果，特徴的な語を 3 語用いた場合の方が 2 語を用いた場合よりも精度が良かったことから，以下の表 4 に示す特徴語を使用して，ユーザの「職業」を推定することにした．

表 4. 各職業における特徴語

社会人	仕事	会社	休み
学生	バイト	学校	テスト
主婦	子供	ママ	旦那

B. 職業推定実験・結果

本研究では，ユーザの「職業」をあらかじめ調査したユーザ 20 人と「職業」が Twitter のプロフィールに記載されているユーザ 10 人の合計 30 人のユーザを実験対象にした．そのため後者の 10 人には Twitter プロフィールに記載されている職業と一致すれば正しく推定されたことにした．また， 30 人の実験対象者は，各職業から 10 人ずつを対象にした．加えて，こちらもツイート数に応じて推定精度に違いが出ることが考えられたため，ツイート数を分けて表示することにした．以下の表 5 に「職業」の推定結果を示す．

表 5.「職業」推定結果

	正			誤		
ツイート数	社	学	主	社	学	主
0~1000	2	3	3	1	0	1
1001~2000	4	1	2	2	0	1
2001~3200	1	6	2	0	0	1
各合計	7	10	7	3	0	3
合計	24(80%)			6		

この結果からある程度推定精度がよく推定された．しかし，推定を誤ったユーザの中には，全ツイートから推定をおこなったため，誤って過去の「職業」が推定された．そこで，「期間を考慮した職業推定手法」を提案する．「期間を考慮した職業推定手法」では誤って過去の「職業」を推定しないようにするため，少ないツイート数から推定をおこなうことにした．しかし，抽出するツイート数が逆に少なすぎることにより推定がおこわれないという問題点をあらかじめ想定したため，最近 1 カ月と 2 カ月をそれぞれ 150 ツイートと 300 ツイート(1 日 5 ツイート×1 月 30 日と算出 [2])と定め，比較実験をおこなった．以下の表 6 と表 7 にそれぞれの推定結果を示す．尚，実験対象にしたユーザは同じユーザである．

表 6. 150 ツイートから「職業」推定結果

	正			誤		
ツイート数	社	学	主	社	学	主
150	8	7	6	2	3	4
合計	21(70%)			9		

表 7. 300 ツイートから「職業」推定結果

	正			誤		
ツイート数	社	学	主	社	学	主
300	9	10	8	1	0	2
合計	27(90%)			3		

以上の結果から，抽出するツイート数が少なすぎる場合逆に推定精度が低下してしまうことが分かった．しかし，300 ツイートで推定をおこなった場合，推定精度が向上されたことが確認された．よって，本研究では職業を推定する場合，最近の 300 ツイートから推定を行うことにした．

VI. 興味推定

この章ではまずユーザの「興味」を推定するために用いた 3 つの手法について詳しく述べる．次に，提案手法では興味の抽出が可能か事前実験や興味の抽出精度を調べるために実験をおこなった．そして最後に推定されたユーザの「興味」についてアンケートをおこなった．

A. 3 つの興味推定手法

「単語の出現頻度による興味推定手法」という手法は，名詞の出現頻度の高い語を特徴的な語とした．具体的には，名詞の出現頻度を調べるために，推定したいユーザの全ツイートを 4 分割に分け抽出をおこない，4 分割されたツイートそれぞれ形態素解析をおこなう．そして，「興味」なりそうな名詞の出現頻度を調べ，特徴的に表れた名詞をユーザの「興味」として推定をおこなう．

「Jaccard 係数を用いた興味推定手法」という手法は，「ソーシャルブックマークを基にした Twitter ユーザの興味語抽出・推薦手法の提案と評価」という論文を参考に考案した．この手法は，用意されたユーザの「興味」ジャンルである“食べ物，スポーツ，ファッション”に関連深い名詞を 30000 ツイートから形態素解析し，抽出する．抽出されたジャンルに関連深い名詞と類似している単語(ラーメンを食べるというツイートの場合，食べるという動詞など)を KH Coder の関連語検索機能を調べ，興味における特徴語として抽出する．そして，プロフィールを作成するときに複数の興味における特徴語と類似している名詞をユーザの「興味」あるものとして抽出する．本研究では「Jaccard 係数」によりそれぞれの語を抽出した．Jaccard 係数は，統計学において 2 つの集合体に含まれるモノや語など類似しているものを推定するために使用される係数である．

「興味辞書を用いた興味推定手法」は，呉らの手法を参考に考案した手法である．この手法は，まず「Jaccard 係数を用いた興味推定手法」と同様に 30000 ツイートを形態素解析し，各ジャンルに関連深い名詞を抽出する．その後，抽出されたジャンルに関連深い名詞と複数のウェブサイトで紹介されているジャンルに関連深い名詞を組み合わせて，ジャンルごとに興味辞書というものを作成する．そして，プロフィール作成時に興味辞書に登録されているジャンルに関連深い名詞とユーザに含まれる「興味」とのマッチングにより出現頻度の高い名詞をユーザの「興味」があるものとして推定した．尚，本研究では複数の試行の結果，他の意味の語と誤ってマッチングした語は排除した．結果的に，興味辞書にはそれぞれ食べ物 246 語，スポーツ 56 語，ファッション 46 語の登録をおこなった．

以上のような 3 つの手法からユーザの「興味」を抽出できるか試みた．本研究ではまず，「Jaccard 係数を用いた興味推定手法」と「単語の出現頻度による興味推定手法」を用いてユーザの「興味」を抽出できるか実験をおこなった．結果「Jaccard 係数を用いた興味推定手法」はユーザの「興味」を抽出することが可能であることが確認された．しかし，「単語の出現頻度による興味推定手法」では，ユーザの「興味」を抽出することができなかった．理由としては，Twitter においては 1 人のユーザから抽出できるツイート数が限られているため，各名詞の出現頻度が少なかったことから，出現頻度に特徴が見つからなかったことが原因であると考える．

B. 興味抽出実験・結果

「Jaccard 係数を用いた興味推定手法」と「興味辞書を用いた興味推定手法」の抽出精度について述べる．抽出精度として「Jaccard 係数を用いた興味推定手法」の場合は，Jaccard 係数の最も高い名詞が各ジャンルに全く関係ない名詞であれば誤りとし，「興味辞書を用いた興味推定手法」の場合は，各ジャンルに関連深い名詞と興味辞書に登録されている名詞とマッチングされなければ誤りとした．以下の表 8 と表 9 にそれぞれの手法における抽出精度について示す．

表 8. Jaccard 係数を用いた興味推定手法による抽出精度

	正	誤
食事	11(55%)	9
スポーツ	5(25%)	15
ファッション	12(60%)	8

表 9. 興味辞書を用いた興味推定手法による抽出精度

	正	誤
食事	20(100%)	0
スポーツ	20(100%)	0
ファッション	19(95%)	1

結果，「Jaccard 係数を用いた興味推定手法」抽出精度があまり良くなかったことから，本研究では「興味辞書を用いた興味推定手法」によりユーザの「興味」あるものを推定することにした．

C. 興味推定の評価

抽出されたユーザの「興味」あるものが本当に「興味」のあるものかを判断するためにアンケート調査(各 5 段階評価であり 5 が一番いいものとする) を 10 人のユーザを対象に実施することで推定手法を評価した．結果を以下の表 10 に示す．尚，本研究では「興味」あるものが 1 つ以上あると考えたことから，本研究では 2 つ記載し，2 つの「興味」から総合的に評価をおこなってもらった．

表 10.「興味」推定のアンケート結果

	満足度					
評価点数	1	2	3	4	5	平均
食べ物	0	0	1	4	5	4.4
スポーツ	0	0	0	2	8	4.8
ファッション	0	2	2	2	4	3.8

この結果からある程度高評価をえることができた．しかし，ファッションにおいては興味がなかったことや興味のないものを抽出したため低い評価になったと考える．

VII. プロフィールの評価

この章ではそれぞれの推定手法により作成されたプロフィールの評価をアンケート形式でおこなう．以下の図 4 にプロフィール例を示し，プロフィールが作成される流れを述べた後，評価結果を示す．

名前	略	
関連地	神奈川県	
職業	学生	
興味		
食べ物	ラーメン	アイス
スポーツ	テニス	サッカー
ファッション	シャツ	リュック

図 4. 作成されたプロフィール例

このプロフィールはプログラミング言語「Java」を使用して，作成されたアプリケーションにより自動作成されたものである．作成までの流れとしては，まず作成されたプロフィールを保存するフォルダーを指定する．そして，プロフィールを作成したいユーザの“Twitter ID”と“抽出したいツイート数”を入力することで自動的に上記の図のようなプロフィールが作成される．

そして最後に作成されたプロフィールに満足したかを 10 人のユーザに「興味」アンケートと同じく 5 点評価で調査おこなった．以下の表 11 に結果を示す．

表 11. プロフィールの満足度アンケートの結果

	満足度					
評価点数	1	2	3	4	5	平均
プロフィール評価	0	0	2	2	6	4.4

この結果からユーザの満足が得られたプロフィールを作成することができたと考える．しかし，中にはユーザの「興味」の種類が少ないため不満の声があったため，“音楽，漫画”などの「興味」の種類を増やすことを今後の課題とする．

VIII. まとめ

本研究では，あらかじめプロフィールを利用して，人々が繋がるときを想定したため，特に必要なユーザ属性「関連地」「職業」とユーザの「興味」3 点を推定し，記載した．先行研究の手法では，本研究で推定したいユーザ属性を高精度に推定をおこなえないことから，本研究では，様々な手法を試みた．結果としては，ユーザの「関連地」と「職業」を高精度に推定することができ実用的な値を出すことができた．特に「関連地」においては，「居住地」のみの推定において 80%の精度であったことから，他の目的に利用するが可能であることが分かった．また，推定したいユーザの“Twitter ID”と“抽出したいツイート数”を入れるだけで自動的にプロフィールを作成できることから本研究の目的であった自動化を実現することができた．しかし，ユーザの「職業」においては，高精度に推定をすることが目的であったため“社会人，学生，主婦”の 3 種類しか推定をできていない．また，ユーザの「興味」においては，アンケート調査である程度ユーザから満足を得ることができたため，ユーザの「興味」あるものを推定することができたと考える．しかし，本研究では，ユーザの「興味」をあらかじめ限定したことで，ユーザから不満の声が出た．そのためユーザの「興味」のジャンルを増やすことが今後の課題である．また，ユーザの「興味」を興味辞書とのマッチングにより出現頻度の多い語をユーザの「興味」として抽出したが，他の意味の語と誤ってマッチングする語は排除したため，ユーザの「興味」になることができないという問題点も今後の課題である．加えて，ジャンルに関連深い名詞の出現頻度からユーザの「興味」あるジャンルを推定し，「興味」あるジャンルにだけを記載するようにしたい．また，「Jaccard 係数を用いた興味推定手法」を提案したが，抽出精度が悪かったことから，最終的にこの手法を用いなかったが，興味辞書に登録されていないユーザの「興味」を抽出することができるため今後の対策次第では有用な手法であると考える．

REFERENCE

[1] 小林亜令, 岩本健嗣, 西山智, “釈迦：携帯電話を用いたユーザ移動状態推定・共有方式”, 情報処理学会 研究報告, 2008-MBL-45, 2008.

[2] 2013 年 SNS 利用動向に関する調査：レポート | ICT 総研市場調査・マーケティングカンパニー, http://www.ictr.co.jp/report/20130530000039.html

[3] 長島直樹, “ソーシャルメディアに表明される声の偏り”, 富士通総研(FRI)経済研究所研究レポート, No.390, 2012.

[4] S. Gauch, M. Speretta, A. Chandramouli, A. Micarelli, “User Profiles for Personalized Information Access”, Lecture Notes in Computer Science 4321, Berlin, pp.54-89, 2007.

[5] 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, Vol. 2, No. 1, 東京, pp.82-93, 2012.

[6] 平野徹, 牧野俊郎, 松尾義博, “Markov Logic を用いたテキストからのユーザ属性推定”, The 27th Annual Conference of the Japanese Society for Artificial Intelligence, JSAI 2013, Toyama, 2013.

[7] 榊剛士, 松尾豊, “ソーシャルメディアからの人物目撃情報抽出システムの試作”, The 25th Annual Conference of the Japanese Society for Artificial Intelligence, JSAI 2011, Morioka, 2011.

[8] 川原大輔, 黒橋禎夫, “要言と直前の格要素の組を単位とする格フレームの自動構築”, 自然言語処理, Vol.9, No. 1, 2002.

[9] 呉勇, 山田祥, 岸本陽次郎, “名詞頻度を使った分類用辞書の構築と評価”, 電子情報通信学会論文誌, Vol. J84-D-I No.2, pp.213-221, 2001.

[10] 小川泰弘, 中村誠, 外山勝彦, “法律文中における単語出現頻度の変化”, 名古屋大学法政論集, Vol.250, pp.543-555, 2013.

[11] 斎藤準樹, 湯川高志, “ソーシャルブックマークを基にした Twitter ユーザの興味語抽出・推薦手法の提案と評価”, 情報処理学会研究報告 情報学基礎研究会報告, pp.1-8, 2011.

[12] 荒川豊, 田頭茂明, 福田晃, “Twitter におけるコンテキストと単語の相関関係分析”, 情報処理学会研究報告 モバイルコンピューティングとユビキタス通信研究会報告, pp.1-7, 2010.

[13] 政府の統計窓口 都道府県・市区町村別統計表, http://www.e-stat.go.jp/SG1/estat/List.do?bid=000001037709

[14] News ポストセブン, http://www.news-postseven.com/archives/20140106_234566.html