ツイッター分析による北海道の盛り上がりリアルタイム抽出

北海道大学大学院情報科学研究科教授　　鈴木　恵二
北海道大学大学院情報科学研究科准教授　　川村　秀憲
北海道大学大学院情報科学研究科学術研究員　　山下　晃弘

第 I 章　はじめに

近年，技術革新とも捉えることができるほど情報通信分野が大きく整備され，その恩恵により，個々のユーザが自ら情報を発信し，またその情報が不特定多数のユーザ間をリアルタイムに経由・閲覧できるようになった．現在ではその情報基盤上において，コミュニティの生成を可能にするソーシャルメディアが注目を浴び，その活用が社会影響を及ぼすまでになっている．ソーシャルメディアは，ユーザが積極的に参加し，双方のコミュニケーションを主要価値と見なすサービスのことで，我が国においては Twitter[1]をはじめとして，mixi[2]，facebook[3]など多岐に渡って利用されている．

本研究では，北海道地域の在住者が，日常生活においてどのような事象に興味を持ち，何を話題とし，口コミ伝搬がどのようなブームを巻き起こしているのかを観測し、データ化することを目的とする．これによって，北海道の観光業者やサービス業者が新たな商品を開発し，顧客に対する戦略的な施策を実現するための IT サービス実現に向けた基礎技術を開発する．本研究では，数あるインターネットサービスの中でも、リアルタイム性が高く，ユーザ数も飛躍的に増加している Twitter に着目し，そこで交わされる様々な情報を分析することで，北海道地域の盛り上がりを観測する手法を検討した．特に Twitter 上で交わされる口コミや，利用者が公開する属性データに基づいて，ユーザのプロファイリングを推定する観点から分析を実施し，どのような話題がどのような人によって交わされているのかを明らかにする要素技術を開発した．

Twitter は 140 文字以内という比較的短い文字制限のある文章を手軽に投稿できるため，隙間時間に自分の近状や意見，主張などを投稿できる軽量なアプリケーションとして人気を呼んでいる．加えて，Twitter ではブログや電子メールなどと異なり，情報がほぼリアルタイムに送受信できること，またその情報が多くのユーザに通知されることから，ユーザの行動履歴を時系列で追跡できるという特徴がある．現に，2011 年 3 月 11 日に我が国において発生した東日本大震災の際には被災地状況の確認や安否情報の取得において大いに活用された．

現在では，Twitter のアクティブユーザは全世界において 1 億人以上いると言われてお

[1] Twitter: http://twitter.com
[2] mixi: http://mixi.jp
[3] facebook: http://www.facebook.com

り[4]，我が国における Twitter の利用者数も年々増加している．図 1 は 2011 年 7 月 11 日時点での我が国における都道府県別 Twitter 利用者数の内訳である．ほぼ全ての都道府県において Twitter の利用者数が 1% 以上となっており，日本全体を通して普及が進んでいることが確認できる．特に東京都，京都府など主要都市においては，利用している人の割合が多く，情報が多数飛び交う大都市地域においては重宝されるアプリケーションとなっていることが伺える．

また，図 2 は都道府県別人口と都道府県別 Twitter 利用者数の対応を示したグラフである．図 2 から得られる相関係数は 0.899 となっており，人口に比例して Twitter の利用者数が増えているのが見て取れる．

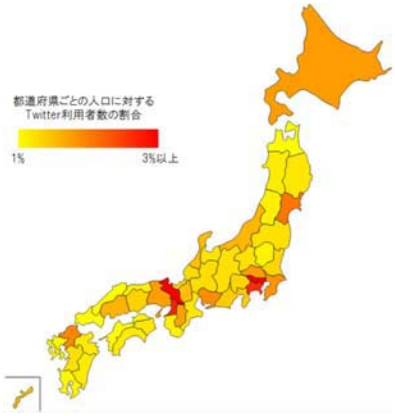

図 1　我が国における都道府県別 Twitter 利用者数の内訳

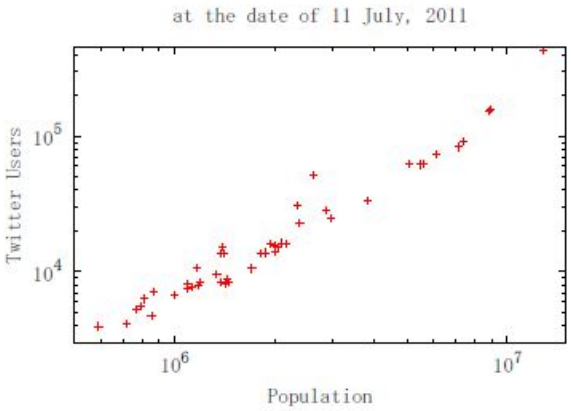

図 2　都道府県別の人口と Twitter 利用者数の対応グラフ

これらを踏まえると，Twitter を用いることで，我々がいつ・どこで・どのように行動し，属性を持ち，興味の傾向があるか認識できる可能性があり，社会の動きをよりよく反映したセンシングツールとしての特性を持つものとして応用できる [1]．さらに，Twitter のよう

[4] http://blog.jp.twitter.com/2011/09/1.html

なソーシャルメディアが登場したことで，商品に対するユーザの心理変容や口コミのトレンド把握，キャンペーンの効果測定など，企業におけるマーケティングの傾聴戦略などへの活用も進んでおり，これらはソーシャルリスニングと呼ばれている[2]．実際，特定のユーザにおけるツイートを人目で通覧することで，ある程度そのユーザが持つプロファイルの推定が可能である．

このような状況を考慮すると，Twitter という情報のクロスプラットフォームにおいてユーザの行動を観察したりプロファイルを推定することは，社会を見つめ有用な情報を発掘したり，特定のプロファイルを有するユーザに対し効果的マーケティングを行うことで，世の中を活性化させる潜在的な要素を持ち合わせていると言えるだろう．

社会的インタラクションを通じて情報が普及されていくソーシャルメディアが注目を浴びている一方，『自ら必要な情報をどのように取得していくか』という時代から，『大量に流れ得る情報をいかに選択し扱うか』といった時代に変化している．これに起因し，莫大な情報が流れるソーシャルメディアを扱うには従来行われてきた情報処理の方法とは全く異なる仕組みが必要になってくる．また，ソーシャルメディアを媒介して流れるテキスト情報は，論文や新聞など既存の構造化された文書とは異なり，文法にとらわれない氾濫した記述形式となっている．そのため，従来のテキストマイニング手法は上手く動作しないという問題点も存在する．このことから，Twitter などに投稿される形式性の低い文からユーザのプロファイルの推定を行う研究は現在まであまり多くなされておらず，推定対象となるプロファイルの粒度及び種類もかなり限定されているのが現状である．

以上の背景を踏まえ，本研究では，新たなソーシャルメディアの一つであるTwitter が世の中に存在する重要な情報発掘に利用可能なツールであることを見越し，有用な情報の一つであるユーザのプロファイルの抽出を試みた．プロファイルの抽出にあたり，本研究ではプロファイルを特徴付けるキーワードを抽出し，API を用いて推定対象となるユーザのメタ情報およびツイート情報から得られるキーワードと類似度を計算するアプローチをとることにより，プロファイルの保有確率を算出した．

Twitter 上におけるユーザのプロファイルが一定制度で推定可能となれば，『あるプロファイルを持つユーザがどのような関心を寄せているのか対象を知りたい』といった調査や，『ある対象に関心を寄せているユーザはどのようなプロファイルを有しているか知りたい』といったマーケティング及び地域振興策の策定に向けた事前調査への応用が見込める．

加えて，ユーザのプロファイル推定手法が確立できれば，膨大な情報が蓄積されたTwitter から有用な情報を発掘する技術として，思わぬ事例の発見に役立つものと予見できる．

第Ⅱ章　Twitter

第１節　Twitterの概要と特徴

Twitter は2006 年7月にObvious 社(現在のTwitter 社) が開始したマイクロブログサービスである．Twitter を利用することで，ユーザは「いつ，どこで，何をしているか」といった自分自身の状況などを140 文字以内という短い文字数で投稿(ツイート：tweet と称

される）することができる．

近年Twitter が認知されるにつれ，我々の情報収集のスタイルも変化しており，以前は新聞やテレビ，Web といった情報に対して，受け手側から自ら情報を受け取りにいく形態が主流であった．ソーシャルメディアの発達により，従来の情報源がリアルタイムに配信されることに加え，自ら取りにいっていた情報が自ずと手に入る環境が構築された．そのため，現在では多量の情報から必要とする情報を取捨選択するスタイルが確立されている．

また，Twitter は実社会の現象と非常に深いかかわりがあり，多くの局面で活用されている．2008 年に開催されたアメリカの大統領選挙にてバラク・オバマ上院議員がTwitter と web サイトを利用することで，支持者同士が容易に集まる環境を構築した結果，大統領選挙において歴史的な勝利を果たしたという事実は，代表的な例と言えるだろう[5]．加えて，2011 年1 月25 日にエジプトで発生した大規模な反政府デモにおいても，住民らがデモを呼びかけるツールとしてFacebook やTwitter を利用された．これにより，エジプトのムバラク大統領の独裁政権に終止符が打たれたという事実は比較的新しい[6]．2011 年3 月11 日我が国で発生した東日本大地震においても，被害の状況把握や避難所情報の確認の他，電話がつながりにくいという状況下での安否確認のために広く利用された[7]．

このように，実世界において社会的に密接した情報のやりとりがTwitter を含むソーシャルメディアを通して行われており，ソーシャルメディアはリアルタイムに状態を観測する目的で多くの活用が期待されている．また，Twitter が持つ特徴として以下のようなものが挙げられる．

- 投稿情報へのアクセスが容易
- リアルタイムな情報の発信や閲覧が可能
- トピックスの種類が豊富
- ネットワーク構造が非対称
- 空間情報が利用可能
- API の利用が可能
- ユーザプロファイルの同定可能性と利用

第２節　Twitterに関する関連研究

前述のとおり，Twitter はソーシャルメディアとして大いに注目を浴びている．また，学術的な観点においてもTwitter が次世代の情報基盤になり得ると期待されている．具体的にはTwitter を次世代の情報共有プラットフォームとして確立するために，Twitter自体がどのような特性が持ち合わせているのか分析・調査する研究と，ソーシャルセンサとしてどのように活用されていくのかを探る研究とに大きく分けられる．図3 はTwitter の基礎特性分析とソーシャルメディアとしての活用の観点から，関連研究を瞰下した図である．

本研究では，投稿されるツイートの特性やユーザが持つ特徴，記載している情報の傾向

[5] http://www.guardian.co.uk/technology/2008/nov/07/barackobama-uselections2008
[6] http://www.guardian.co.uk/world/2011/feb/25/twitter-facebook-uprisings-arab-libya
[7] http://tr.twipple.jp/info/bunseki/20110427.html

などを効率良く把握するため，Twitter API とCakePHP を用いてツイッターのつぶやき情報を効率よく収集し，分析するための専用Web サイトを構築した．詳細は後述するが，このサイトを利用することで，ユーザが投稿するツイートの内容や全体の投稿時間帯，利用するクライアントソフト割合，特定キーワードを含むツイートの投稿時間の比較，フォローしているユーザの紹介文に記載されているキーワードの頻度など幅広い分析を行うことが可能となった．

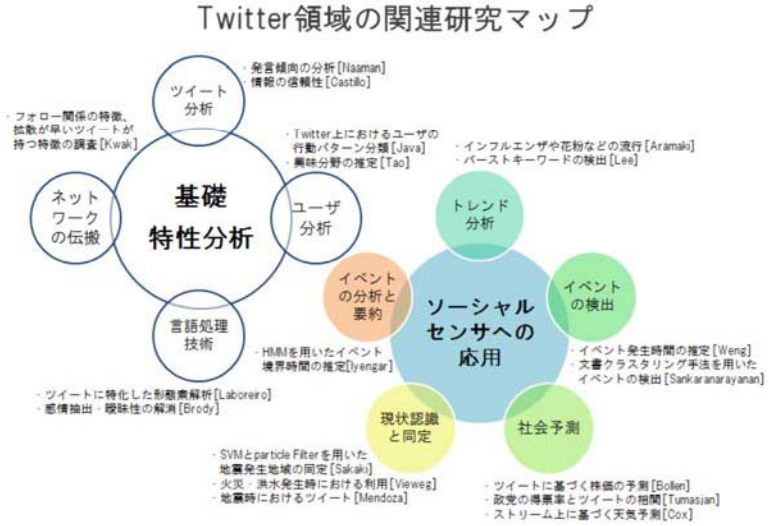

図 3　現在までに取り組まれている Twitter に関する研究の俯瞰図

第Ⅲ章　ツイッターのつぶやき情報のリアルタイム収集システム

図4は本研究で構築した分析用Web サイトの外観を示したものである．分析用Web サイトでは，本研究で収集した任意のユーザ(1,729,073人）の基本的な情報を閲覧することが可能である．ここでは，ツイート数やリツイート数，フォロー数，フォロワー数，最も発言する時間帯と曜日及びロケーションデータと自己紹介欄が表示される．

図5はユーザが投稿したツイートに関する詳細な情報を表示したページの様子である．グラフの描画にはGoogle Chart API[8]を利用しており，ツイートにジオタグを付与するクライアントソフトの利用割合についての内訳や，ツイートの文章の長さから，意見の主張・暇つぶし・日記の代替など，ユーザがTwitterをどのような用途で利用しているかといった点を考慮する際に役立つものと思われる．

図6はユーザがフォローしているユーザの自己紹介文に出現するキーワードの種類と頻度を表示したページである．自分がフォローした相手のツイートは自身のタイムラインに表示されるため，ユーザは自分の興味のある内容をツイートしているユーザや友達，ビジネスパートナーなどが主なフォロー対象になるものと考えられる．すなわち，あるユーザがフォローしているユーザについて分析することにより，ユーザの興味対象やプロファイ

8 Google Chart API: http://code.google.com/apis/chart

ルに寄与する可能性があるため，フォローしているユーザの特徴を全体的に俯瞰したい場合に有効な情報である．

また，構築した分析用Web サイトはユーザの軸とした分析だけではなく，キーワードを軸にした分析も可能となっている．図7 は指定したキーワードを自己紹介文に含んでいるユーザの一覧を表示したページである．

これは，ユーザがあるプロファイルを明示的に有しているか確認したい時に，またはそれらのユーザ一覧を集めたい時に役立つものとなっている．例えば「札幌」とキーワードを入力してページを更新すると，自己紹介文で札幌について言及しているユーザの一覧が表示されるため，札幌を中心に活動しているユーザ集合の確保したり，札幌に興味があるユーザの特徴を全体を通して眺めることが可能である．

本研究においては，次節以降にて性別の推定アルゴリズムを設計する際に，男性と女性のユーザ集合を集める点で活用した．図8 では，ツイート内において指定したキーワードと共起するキーワードの種類と頻度を表している．あるキーワードに関連するキーワードを集めることでテキストマイニングの前処理として行われるクラスタリングや，特異的なキーワードの発見などが期待できる．

図9は，あるキーワードを含むツイートを投稿時刻と共に表示させたものである．図9の例を用いて述べれば，札幌というキーワードがどのようなタイミングでツイートに含み投稿されたか，時系列的に追うことができるため，キーワードの注目や特徴の変化を追うツールとして活用できる．

図4　構築した分析用Web サイト
（Twitter におけるユーザの基本情報）

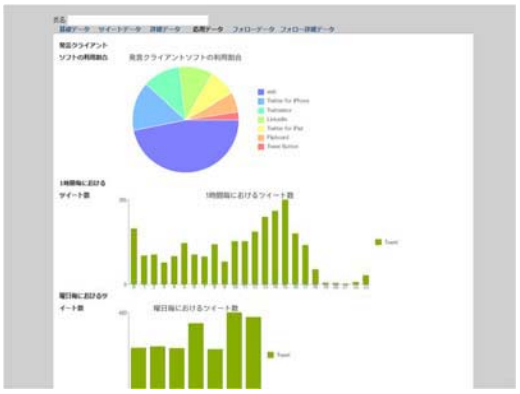

図5　クライアントソフトや
ツイートに関する情報の詳細

図6　フォローユーザの自己紹介文に
出現するキーワードの種類と頻度

図7　あるキーワードを
自己紹介文に含んだユーザー覧

図8　ツイート内においてあるキーワードと同時に出てくるキーワードの種類

図9　あるキーワードを含むツイートー覧

第Ⅳ章　つぶやき情報の分析アルゴリズム

第１節　ユーザプロファイル

前述のとおり，Twitter に関する研究は今まで多く取り組まれてきている一方，サービスを利用するユーザ自体を分析する研究はあまり多くないのが実情である．本研究では，このユーザ自体が持つプロファイルを Twitter 上において推定することを目的とし，ここでは，そのユーザプロファイルについて述べる．

（１）推定定対象とするユーザプロファイル

本節では，推定を試みるユーザのプロファイルを定義する．ユーザプロファイルの定義は研究により様々であり，定義によっては非常に幅広く捉えることもできる．ここで，ユーザ $u \in \mathbf{U}$ のプロファイル P を以下の式のように質問 $q \in \mathbf{Q}$ と回答 $a \in \mathbf{A}$ から決定されるものと仮定する．

$$f : \mathbf{U} \times \mathbf{Q} \rightarrow \mathbf{A}$$
$$P = \{ f \mid u \in \mathbf{U}, q \in \mathbf{Q}, a \in \mathbf{A}, f(u,q) = a \}$$

つまり，質問の数だけプロファイルが定義されるため，その数は無限に等しい．例を挙げれば，『好きな色は何ですか』といった質問は回答が有限個であるものの，回答の数は実に幅広い．また，『好きな女性のタイプはどのような方ですか』といったような質問は，回答するユーザによって全く異なり，予め決められたプロファイルが存在しない例である．このようにプロファイルを推定する際，予め決められた回答が存在しない場合，また回答が有限の種類であったとしても，その全てを定義しなければいけない場合はそもそも主体的な推定が困難である．(このようなプロファイルを本研究では『内装的プロファイル』と呼ぶ)

以上から，プロファイルを推定する際には，予め『年齢』や『性別』などの質問に対し，

『10 代, 20 代，…』，『男性，女性，不明』といった有限の回答のどれに当てはまるかといったアプローチが必要になる．(これらのプロファイルを『外装的プロファイル』と呼ぶ)

本研究においては，外装的プロファイルを前提として，プロファイルを『人間が持つ生理的・社会的な属性』と定義し，下記に示す種類，粒度の推定を目指す．

・性別(男性・女性)
・居住地域(人口の多い市から上位20)
・職業(代表的な種類，学生，主婦)

（２）ユーザプロファイルに関する既存研究

ユーザプロファイルについては既存研究によって定義がまちまちであり，本研究の定義と同じ研究もあれば，ソーシャルメディアの特性をユーザの特性として扱っている研究も多い．既存研究に関する詳細については本稿では割愛するが，具体的な研究例を参考文献[3)4)5)6)7)]に示す．

第２節　利用するデータセット

本研究では，Twitter API を用いてTwitter のユーザ情報，ツイート内容及びメタ情報，フォロー関係及びツイートのジオタグ情報を収集した．収集期間は2011年6 月11 日から2012 年1 月27 日とし，データの詳細内容を表1に示す．

表1　利用するツイートのデータセット

ユーザ数	1,729,073
平均ツイート数	3913
平均フォロー数	203
平均フォロワー数	235
のべツイート数	107,606,168
ジオタグ付きツイート数	279,101(全ツイートの0.26%)

また，Twitterのロケーション設定欄や自己紹介欄において既にプロファイルそのものが記載されている場合も多い．ロケーションの設定者数に関するデータを表2に示す．なお，表2におけるロケーションデータでは県と市町村の該当対象を調べるために，市町村名一覧表として『総務省統計局・製作総括官(統計基準担当)・統計研修所統計に用いる標準地域コード(全国版)』[9]を利用した．また，ロケーションデータに必ずしも正しい県及び市町村名が記入されているとも限らないため，本研究では形態素解析エンジンであるMeCab[10]を利用してロケーション内容を素性に分解し，漢字かなまじり文をローマ字に変換するツール

9 統計局標準地域コード: http://www.stat.go.jp/index/seido/csv/9-5.csv
10 MeCab: http://mecab.sourceforge.net

であるkakasi[11]を利用することで, 標準地域コードと一致するユーザのみをカウントしている．表2から約40% のユーザに関しては市町村レベルまで公開していることがわかる．これより，居住地域の推定実験を行う際の訓練データ及び正解データとして，これらのユーザ情報を活用していく．

表2　データセットにおけるロケーションの設定状況

ロケーション設定者数	1, 262, 752（73. 03%）
県レベルまで記載している人数	771, 913（44. 64%）
市町村レベルまで記載している人数	689, 890（39. 90%）

第３節　ユーザプロファイルの特徴を表す指標の定義

本節では，プロファイル推定に向け，これまで本研究で進めてきたプロファイルの特徴量をキーワードの重み付けの観点から取得するための手段を述べると共に，各々の指標について考察を述べる．これらの指標全体を定義したのち，ユーザのプロファイルを取得する上で最適な手法を選択し，各々のプロファイルに向けたアルゴリズムの設計を行う．

（１）TF-IDF によるキーワードの重みづけ

テキストマイニングの分野においてキーワードを軸としたアプローチをとる場合，各キーワードに重みを付け，任意の文を特徴付けるのが主流とされている．代表的なものとして，キーワードの出現頻度を考慮したTF/IDF法が挙げられ，以下の式で表現される．

$$TF-IDF\ (\omega, d) = TF(\omega, d) \cdot \log(N / DF(\omega))$$

ここで, $TF(\omega, d)$はユーザのツイートd におけるキーワードωの出現頻度, Nは全ユーザ数, $DF\ (\omega)$ はキーワードωを含むツイートを投稿しているユーザ数である．つまり，同じユーザがツイートを投稿する時，ツイート内に何度も出現するキーワードほどそのユーザを特徴付けると考え，その特徴を表すものがTF値である．反対に多くのユーザが投稿するツイートに多く出現するキーワードは, ユーザを差別化できないものとして捉えた指標が$\log(N / DF(\omega))$である．

本研究では，後に示すコサイン類似度を用いた相関値を求める際，キーワードを選択するステップにおいて，ここで定義したTF値を用いている．

（２）偏差値を用いたユーザの特異性

プロファイルを特徴付けるキーワードが算出された際，そのキーワードを持つユーザとどれだけの差があるのかを表現する場合，偏差値を用いて表現することができる．まず，Twitter 上のユーザ集合$\mathbf{U} = \{u_1, u_2, \cdots, u_i\}$として表し，ユーザ$u^*$ がフォローしているユー

[11] kakasi: http://kakasi.namazu.org/

ザ集合を $F_{u*} \subset \mathbf{U}$ と表現する．

加えて，ユーザu^*のツイート集合を

$$\mathrm{Tweet}_{u^*} = \{t_{u^*1}, t_{u^*2}, \ldots, t_{u^*j}\}$$

とする．ここでTwitter上のウェブページにおけるユーザu^*の説明文を$description_{u^*}$，ロケーション情報を$location_{u^*}$とし，ユーザu^*が有するドキュメント情報を

$$\mathbf{D}_{u^*} \in \{description_{u^*}, location_{u^*}\},$$
$$\mathbf{D}_{u^*} \subset \mathrm{Tweet}_{u^*}$$

と定義する．ドキュメント集合$\mathbf{D}$の要素d に含まれるキーワード集合は

$$\mathbf{K}_{d \in \mathbf{D}} = \{k_{d1}, k_{d2}, \cdots, k_{dl}\}$$

と定義できるため，あるユーザu^*がもつドキュメント集合D_{u^*}の要素dにおいて，キーワード集合$\mathbf{K}_{d \in \mathbf{D}u^*}$の中に特定のキーワード$k^*$が含まれているかどうかを判別する二値関数 f は

$$f_{K_{d \in \mathrm{D}_{u^*}}}(k^*) = \begin{cases} 1 \ (k^* \in K_d) \\ 0 \ (k^* \notin K_d) \end{cases}$$

と表現可能である．

これらの定義を踏まえ，あるユーザu^*がフォローしているユーザのドキュメント内に特定のキーワード$k*$が含まれているとすると，フォローユーザ1人あたりにおけるk^*の出現期待値は

$$ev(u^*, k^*) = \frac{\sum_{u \in \mathrm{F}_u^*} f_{K_{d \in D_u}}(k^*)}{n(\mathrm{F}_{u^*})}$$

となる．さらに，全ユーザに関するk^*の出現期待値は

$$avg(k^*) = \frac{\sum_{u_i \in \mathrm{U}} ev(u_i, k^*)}{n(\mathrm{U})}$$

と算出される．以上より，キーワードk^*における出現頻度の標準偏差は

$$\sigma(k^*) = \sqrt{\frac{\sum_{u_i \in \mathrm{U}} (ev(u_i, k^*) - avg(k^*))^2}{n(\mathrm{U})}}$$

と求められることから，ユーザu^*におけるキーワードk^*の偏差値は

$$\Theta(u^*, k^*) = \frac{ev(u^*, k^*) - avg(k^*)}{\sigma(k^*)}$$

と導かれる．

この偏差値は，キーワードk^*に関して，特定のユーザu^*が他のユーザと比べてどれだけ特異的かを表す．あるユーザが投稿したツイートにおいて，出現するキーワードを他のユーザのツイートにも含まれていたとした場合，そのキーワードが地域性の高い特徴を有する，または専門性の高い特徴を有すると，ユーザ同士は同じ特徴を有する可能性がある．それを判断する際に，この指標が有効であると考えられる．

（３）発言傾向に基づくユーザの特異性

ある特定のキーワードk^*を含むツイート集合を$\mathrm{Tweet}(k^*)=\{t_{k^*1}, t_{k^*1}, \cdots, t_{k^*j}\}$とする．ここで，ツイート集合Tweetからユーザ集合**U**を求める関数gは，

$$g: \mathrm{Tweet} \rightarrow \mathbf{U}$$

と表すと，ある特定のキーワードk^*を含むツイートを投稿したユーザの集合$\mathbf{U}(k^*)$は，

$$\mathbf{U}(\mathrm{k}^*) = \mathrm{g}\left(\left(\bigcup_{u_i \in \boldsymbol{U}} \mathrm{Tweet}_{u_i}\right) \cap \mathrm{Tweet}(k^*)\right)$$

となる．さらに，ユーザu_iの居住地域と職業をそれぞれ$location_{ui}$ と$career_{ui}$で表すと，全居住地域と職業を表す要素は

$$\mathbf{L} = \bigcup_{\boldsymbol{u_i} \in \boldsymbol{U}} location_{u_i}$$

と定義できる．ここで，ユーザが持つプロファイルe_{u^*}が特定のプロファイルe^*に一致するならば1，一致しなければ0 を返す二値関数

$$I_{e^*} = \begin{cases} 1\ (e_{u^*} = e^*) \\ 0\ (e_{u^*} \neq e^*) \end{cases}$$

を定義すれば，ある特定のキーワードk^*を含むツイートを投稿したユーザの集合$\mathbf{U}(k^*)$において，最も多くの割合を占める居住地域$location*$ と職業$career*$は，

$$location^* = \underset{location_j \in \mathbf{L}}{\arg\max} \sum_{u_i \in \mathbf{U}(k^*)} I_{location_j}(location_{u_i})$$

$$career^* = \arg\max_{career_j \in \mathbf{C}} \sum_{u_i \in \mathbf{U}(k^*)} I_{career_j}(career_{u_i})$$

と求められる．これは，あるキーワードk^*を含むツイートを投稿したユーザのプロファイルを集計し，その中から最大数となるものを推定対象となるユーザのプロファイルとして割り当てることを表している．図10はこのアルゴリズムの流れについて説明したものである．

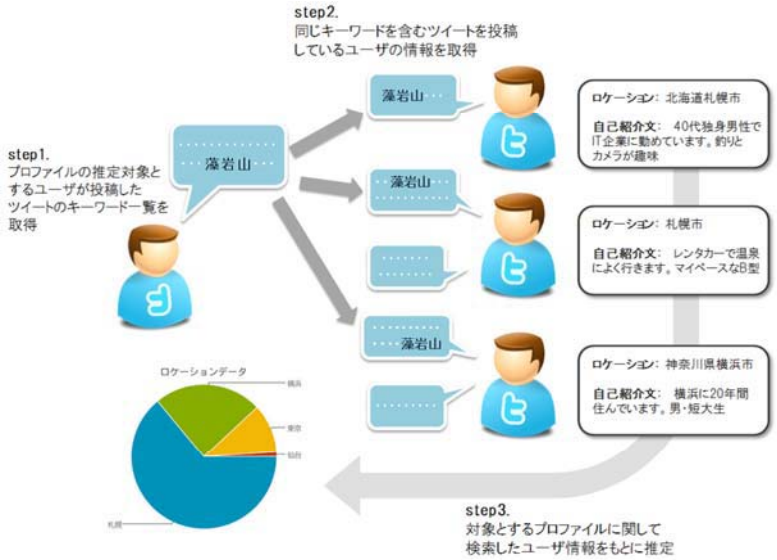

図10　発言傾向に基づくユーザの特異性を計算するアルゴリズムの概略図

（４）情報量を用いたキーワードの特異性指標

あるキーワードがプロファイルの粒度において重み付けされ，それが確率分布として表現されている時，そのキーワードを含むツイートを投稿した時どれだけ，該当するプロファイルを有する確率があがるかを判断する際には，カルバック・ライブラ・ダイバージェンスを用いて表現できる．

まず，全ユーザ集合$\mathbf{U}$の離散確率分布を$P(X)$，あるキーワードk^*を含むツイートを投稿しているユーザ集合$\mathbf{U}'$の離散確率分布を$Q(X) = P(X|k*)$とする．ここで，$\mathbf{U}$の自己情報量は$-logP(X)$ であり，$\mathbf{U}'$の自己情報量は$-logQ(X)$である．もし，あるキーワードk^*を含むツイートは特定のユーザに限らず投稿され，すなわち，$\mathbf{U}'$が$\mathbf{U}$と同じ分布になるとすれば，k^*を含むツイートから得られる情報量は0 である．一方で$\mathbf{U}$と$\mathbf{U}'$が互いに異なる分布であるとすれば，キーワードk^*を含むツイートはある特定のユーザにおいて投稿される傾向があるので，その差分となる情報量は

$$\big(-\log P(X)\big) - \big(-logQ(X)\big) = log\frac{Q(X)}{P(X)}$$

と算出される．xはXに従って変化することから，この値の平均値を取ると，

$$D_{KL}(Q||P) = \sum_{x \in X} Q(x) log \frac{Q(x)}{P(x)}$$

となる．D_{KL} はカルバック・ライブラ・ダイバージェンスと呼ばれ，U に対する U' の情報利得を意味する．

（５）コサイン類似度を用いた相関値

本研究では，コサイン類似度を用いて性別に関するプロファイルの推定を行う．なお，コサイン類似度の算出に当たっては，土井らの研究[8]の手法を用いることとし，詳細な算出手法についての記述は割愛する．

第４節　ユーザプロファイル抽出アルゴリズムの設計

本節では，ユーザプロファイルの中から性別を例にとり，性別を推定するアルゴリズムについて述べる．ここでは，コサイン類似度を用いた相関値により，任意のユーザが男性・女性を推定する．グループは男性と女性の二種類とし，データベースより予め目視で性別の区別がつくユーザを男女各500名ずつ用意する．男性・女性を推定する際には，ユーザのプロファイルやツイートを目視し，ユーザが直接自身の性別を記載している，または性別を明らかに判断できる記述があるものに関して性別を区別してグルーピングを行う．(表3参照)

表3　性別を区別する際に用いた自己紹介文の例

性別	自己紹介文の例（抜粋）
男	日本語ラップが好きな17 才の男。…
男	立川を愛してやまない男です。大学3 年生…
男	フットベースを指導しているオレにフォロワーお願いします。
女	お腹の弱い16 才、高2、女。妄想と音楽が主食…
女	名古屋で1 児の母やってます。手芸と着物がライフワークです。
女	… 只今、妊娠中です(´ー`*)))) 今年の冬に生まれます！宜しくー

キーワードについては，まず始めに男女全ユーザ1000 人が発言したツイートを素性に分解する．一般的ツイートを素性に分解する際に用いる形態素解析エンジンとして，MeCabやChaSen[12]，Juman[13]などが挙げられるが，本研究では辞書やコーパスに依存しない汎用的な設計がなされている点で広く一般的に用いられているMeCab を使用する．

素性に分解した後，素性ごとに出現頻度$P(feature)$と出現回数を求める．同様に男性ユーザ500 名のみのツイート集合から出現頻度$P(feature|male)$と出現回数を算出する．ここで，全ユーザのツイート数に対して男性ユーザのツイート数の割合を$P(male)$とすると，以下のベイズの定理

[12] ChaSen: http://chasen-legacy.sourceforge.jp
[13] Juman: http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN

$$\mathrm{P}(male|feature) = \frac{P(feature|male)P(male)}{P(feature)}$$

から，ある素性が性別の違いによりどの程度出現れやすい，すなわち性別を判断する際の特徴指標として利用できる．

ベイズの定理から導かれる*P*(*male*|*feature*)は性別の推定という点で二値問題であり，値が1に近ければ男性，0 に近ければ女性が発言するツイートによく含まれる素性と解釈できる．これより，*P*(*male*|*feature*)が0.6 以上また0.4 以下の値をとり，かつ男女両方で出現する回数が多い素性に注目し，性別を推定する際に男女各々を特徴付けるキーワードとして以下のものをピックアップした．選んだキーワードを表4に示す．また，表5 に*P*(*male*|*feature*)の範囲ごとにおけるキーワードの数について示す．表5を見て見ると，男性より女性を特徴付けるキーワードが圧倒的に多いことが伺える．実際，男性ユーザは男女共に一般的に使われるキーワードを多用する傾向がある一方，女性は男性が使わないような語や顔文字を多くツイートに含んでいる傾向が見られた．表4で選んだキーワードを用いてキーワードベクトルを作成し，コサイン距離を用いて男女の推定を行うが，このアルゴリズムの性能については適合率(precision)と再現率(recall)の両者を用いてF1指標と呼ばれる値を算出し評価する．

表4　性別を区別する際に用いた自己紹介文の例

キーワード	$P(male\|feature)$	出現人数（1000人あたり）
俺	0.836	911
オレ	0.757	409
おれ	0.757	497
バイク	0.710	428
僕	0.692	845
お前	0.634	792
ギター	0.634	429
サッカー	0.629	622
ぼく	0.625	463
彼氏	0.337	655
ランチ	0.213	533
(*＾＾*)	0.187	308
夫	0.186	394
化粧	0.183	542
私	0.161	980
赤ちゃん	0.147	458
肌	0.139	622
わたし	0.109	677
あたし	0.104	466
旦那	0.037	549

表5　各重みの範囲におけるキーワード数と頻度

$P(male \mid feature)$の範囲	範囲に含まれるキーワード
$0.9 > P$	2
$0.9 \leqq P < 0.8$	6
$0.8 \leqq P < 0.7$	17
$0.7 \leqq P < 0.6$	71
$0.4 \leqq P < 0.3$	435
$0.3 \leqq P < 0.2$	109
$0.2 \leqq P < 0.1$	71
$0.1 \leqq P$	42

カテゴリに分類された全ての数をN，正しく推定された数をR とすると，適合率は

$$precision = R / N$$

で表され，推定結果として得られたものの中にどれだけ推定に適合した数が存在するかという正確性の指標と捉えられる．

また，本来カテゴリに属する全ての数をC とすると，再現率は

$$recall = R / C$$

によって求められ，推定結果として適合した数のうちどれだけカテゴリに含まれているものを推定できたかという網羅性の指標である．

本来，適合率と再現率はトレードオフの関係であることから，一方があがれば一方が下がるため，両者の調和平均である$F1$ 指標が評価としてよく用いられ，

$$F_1 measure = \frac{2 * precision \cdot recall}{precision + recall}$$

で与えられる．これはR をN とC の相加平均で割ったものに相当し，$F1$ 指標が高ければ，性能が良いことを意味するものとして用いることができる．第Ⅴ章の実験においては，男性と女性それぞれにおける適合率と再現率，またそれらから算出される$F1$ 指標を用いてアルゴリズムの性能を検証する．

第Ⅴ章　実証実験

第Ⅳ章で定義した指標及びアルゴリズムを用いて，プロファイルの推定にむけた実験の設定と結果に対する考察を行う．男女それぞれ100 名で1 つのクラスタとし，男女それぞれ3 つのクラスタを組み合わせた300 名を訓練事例としてグループベクトルの算出に利用し，残りの2 つを組合せた200 名を実際にプロファイルを推定できるかどうかのテスト事例として利用する．また，本研究では，クロスバリデーション10）を用いて，モデルの汎用性を検証する．

ここで，クロスバリデーションとは訓練事例をK個に分割することで，K-n個のデータセットをモデル推定に使用し，残りのn個のデータセットをモデルの評価にテスト事例として用いる．モデル推定に用いたデータセットに含まれないデータによってモデルの評価を行うことで，認識時に近い形でモデルの評価を行うことが可能となる．今回のクラスタは男女各々5 つずつあるため，クラスタの組合せを変えることで得られる5C2 = 10通りの訓練・テスト事例について検証可能である．これらの実験設定をもとに，男性・女性それぞれにおける適合率と再現率は表6のようになった. 表6から, 適合率は男性･女性とも80% を越えており，コサイン類似度のアルゴリズムを用いることである程度の性別の推定が行えることが伺える．また，標準偏差も2% 以内となっており，精度のばらつきは少ない．再現率に関しては，女性の平均値が高くなっており，本来女性であるユーザが女性と判別される割合は非常に大きいという結果になった．F1 指標に関しても男性・女性の両方において0.85 以上の値となっており，性別の推定においてはコサイン類似度のアルゴリズムが有効であることと捉えることができる．

表6　各重みの範囲におけるキーワード数

検証パターン	適合率（precision）		再現率（recall）	
	男性	女性	男性	女性
1	91.6%	79.5%	76%	93%
2	95.9%	84.3%	82%	96.5%
3	93.6%	82.5%	80%	94.5%
4	95.2%	82.4%	79.5%	96%
5	93.5%	81.5%	78.5%	94.5%
6	96.5%	85.5%	83.5%	97.5%
7	93.9%	80.2%	76.5%	95%
8	92%	78.6%	74.5%	93.5%
9	96.3%	81.9%	78.5%	97%
10	96.3%	80.8%	77%	97%
平均値	94.5%	81.7%	78.6%	95.5%
標準偏差	1.7%	2.0%	2.6%	1.5%
F1指標		0.858		0.881

第Ⅵ章　まとめ

本研究では，様々な情報が行き交うネット社会において，どのような人が今何を話題とし，口コミ伝搬がどのようなブームを引き起こすのかを明らかにするための要素技術を開発した．今回は，リアルタイムに膨大な情報がやり取りされているTwitter を対象とし，ある話題がどのような人たちによって発信されているのかを探るための有用な一つの情報と考えられるユーザのプロファイルの抽出を試みた．またその中でも性別・居住地域・職業の推定についてアルゴリズムの設計と推定を目指し，本稿においては性別の推定を実験を行うことで試みた．

本研究の実験結果から，キーワードを特徴量として適切に選択し重み付けし，コサイン距離を用いた類似度を計算するアプローチをとることで，性別においては80% 以上の精度で推定可能であることがわかった．

今後の展望としては，性別だけでなく居住地域や職業，さらにはその他の種類におけるプロファイルの推定についても考え，Twitter を利用するユーザ層を明らかにする．また，その結果から北海道地域における話題の盛り上がりや，ブームのリアルタイムな把握などが可能なシステムに応用し，観光産業やサービス事業者の方々が地域の経済振興に向けた施策や商品開発に応用可能な仕組みに仕上げていく予定である．

参考文献

1) 榊剛史，松尾豊: ソーシャルセンサとしてのTwitter -ソーシャルセンサは物理センサを凌駕するか?-: 人工知能学会誌，pages 67-74，Vol.27，No.1，January 2012. 社団法人人工知能学会.

2) K.Crawford: Following You: Disciplines of Listening in Social Media: Con-tinuum, Journal of Media & Cultural Studies Vol. 23, No. 4, pages 525-535, August 2009.

3) K. Tao, F. Abel, Q. Gao, and G. J. Houben: TUMS: Twitter-based User Modeling Service.: In Proceedings of International Workshop on User Profile Data on the Social Semantic Web (UWeb), co-located with Extended Semantic Web Conference (ESWC), Heraklion, Greece, 2011.

4) Konstan, J., Conejo, R., Marzo, J., Oliver, N. (eds.) User Modeling, Adaption and Personalization, Lecture Notes in Computer Science, vol.6787, pages 1-12, 2011. Springer.

5) C. M. Au Yeung, N. Gibbins and N. Shadbolt: A study of user profile generation from folksonomies: In WWW'08: Social Web and Knowledge Management, Social Web 2008 Workshop at Proceedings of the 17th International World Wide Web Conference, Beijing, China, April 2008.

6) M. Thelwall: Social Networks, Gender, and Friending: An Analysis of MySpace Member Pro_les: Journal of The American Society for Information Science and Technology, pages 1321-1330, 2008.

7) D. Rao, D. Yarowsky, A. Shreevats, and M. Gupta: Classifying latent user attributes in twitter.: In Proceedings of the 2nd international workshop on Search and mining user-generated contents, SMUC'10, pages 37-44, New York, NY, USA, 2010. ACM Press.

8) 土井俊介，吉田由紀，東野豪: RUI-Filtering:利用履歴のアイテムの類似関係を反映した協調フィルタリング方式: 電子情報通信学会データ工学ワークショップ(DEWS2004), I7-6, 2004.

9) 松原望: 入門ベイズ統計- 意志決定の理論と発展: 2008. 東京書籍.
10) C. M. Bishop, 元田浩, 栗田多喜夫, 樋口知之, 松本祐治, 村田昇: パターン認識と機械学習- ベイズ理論による統計的予測, 2006, Springer.
11) R. Feldman, J. Sanger, 辻井潤一, IBM 東京基礎研究所: テキストマイニングハンドブック: 2010. 東京電機大学出版局