Report レポート #01

(財)北海道開発協会平成23年度研究助成サマリー

ツイッターからの情報抽出とその展望

～北海道の盛り上がり観測に向けて～

鈴木　恵二（すずき　けいじ）
北海道大学大学院情報科学研究科教授

1965年北海道壮瞥町生まれ。北海道大学大学院工学研究科精密工学専攻博士後期課程修了。博士（工学）。93年から北海道大学工学部助手、助教授を経て、2000年より公立はこだて未来大学にて助教授、教授となる。08年より現職。人工知能、自律ロボット、観光情報学、サービス工学等の研究に従事。

1　はじめに

近年、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をはじめとして、mixi、facebookなど多岐にわたって利用されている。

本研究では、北海道地域においてどのような事象に興味が生じているのか、何を話題とし、口コミ伝播（でんぱ）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しようとするものである。特に、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。このとき、ある投稿が北海道在住の方からのものなのか、あるいは他地域からの観光客なのか、ビジネスマンなのか、女性か男性か、いくつぐらいの方なのか、発言者の大まかなプロフィールが推定できれば、その投稿の活用も大きく広がることになる。そこで、投稿履歴や、ユーザが公開する属性データに基づいて、ユーザのプロファイルを推定する技術開発についての概略を説明する。

2　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。

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､時間にそって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１億人以上いるといわれており、我が国におけるTwitterの利用者数も年々増加している。2011年７月11日の時点での我が国における都道府県別Twitter利用者数の内訳を調べると、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。また、都道府県別人口と都道府県別Twitter利用者数の対応を調べた結果、その相関係数は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確認された。

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、どのような興味を持つかを調査できる可能性があり、社会の動きを反映したセンシングツールとしての特性を持つものと期待される。さらに、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、ソーシャルリスニングと呼ばれるマーケティングの傾聴戦略などへの活用、研究も盛んに行われるようになった。図１に関連研究の俯瞰図を示す。

一方、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うま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、あまり研究されておらず、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、口コミといった情報をより有効活用するため、基盤情報となるユーザのプロファイルの推定を試みる。すなわち、Twitter上におけるユーザのプロファイルがある程度推定可能となれば、例えば北海道の観光において、単純にどのような話題が盛り上がっているかだけでなく、観光客の年代別、性別ごとの関心事をリアルタイムに把握できるとともに、地元の方々が、旬な情報として発信しているものを区別し、マッチングを図っていくといった応用なども期待できるであろう。

3　ツイートのリアルタイム収集システム

北海道の盛り上がりを観測するとともに、その状況把握をさらに進めるものとしてユーザのプロファイルを推定するという本研究推進のための基盤として、投稿されるツイートやユーザ情報などを収集し、分析す

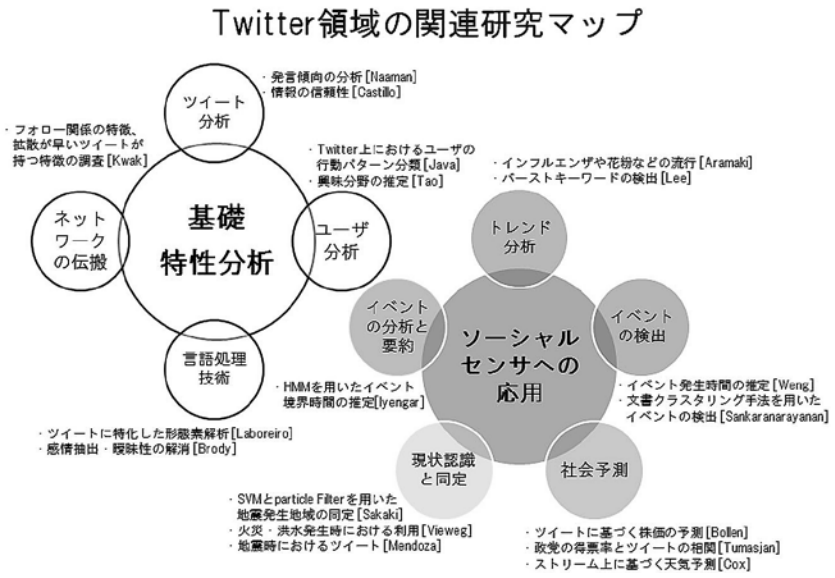

図１　現在までに取り組まれているTwitterに関する研究の俯瞰図

るための専用Webサイトを構築した。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の割合、特定キーワードを含むツイートの投稿時間の比較など幅広い分析を行うことが可能となった。

その機能の一部を紹介する。図2は、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文が表示される。

図3は、ユーザが投稿したツイートに関する情報をグラフ表示したページの様子である。ツイートの文章の長さから、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なっている。

図4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、あるプロファイルを明示的に有している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、札幌に興味があるユーザの特徴を眺めることが可能となる。

図5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。例えば、「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まれたか、時系列で追うことができるため、注目のキーワードやその特徴の変化を追うツールとして活用できる。

4　プロファイル推定

前述の分析サイトを活用することによって、今何について盛り上がっているかなど、話題の抽出や、目視のレベルでどのような人々がその話題に興味を持っているかということが把握可能となっている。しかし、もうひとつ踏み込んだところで、状況分析を行いたい

図2　構築した分析用Webサイト（Twitterにおけるユーザの基本情報）

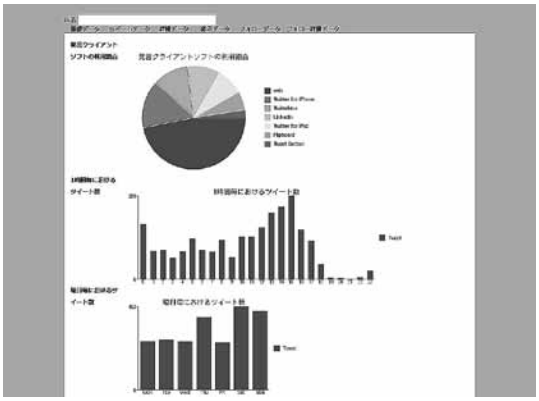

図3　クライアントソフトやツイートに関する情報の詳細

図4　あるキーワードを自己紹介文に含んだユーザー覧

図5　あるキーワードを含むツイート一覧

と思った場合、すなわち、先述の観光の例であれば、地元の方なのか、ビジネスで北海道を訪れている方なのか、日頃からどんなことに興味を持っている方なのか、その背景を踏まえて、ツイートの分析を試みようとすると、そのままでは先に進まないことがわかる。つまり、プロフィールとして明示的に記載しているユーザはごく一部であり、大多数の人々は、そのままではどのようなユーザか不明である。よって、ツイートの傾向等からそのユーザのプロファイルを推定することで、ツイートの背景まで含めた理解が可能となる。この趣旨のもと、本研究では、ユーザのプロファイルを推定する手法の開発を行った。

プロファイルの推定を目指すにあたり、まずはプロファイルとは何かを決めなければならない。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。ここではあるユーザのプロファイルを質問と回答から決定されるものと仮定する。特に、『年齢』や『性別』などの質問に対し、『10代、20代、…』『男性、女性、不明』といった客観的で有限の回答のどれに当てはまるタイプのプロファイルかを推定することとした。具体的には、性別、居住地域、職業とした。

上記プロファイルをその人の過去のツイートやフォローしている人々の自己紹介文等から推定するための基本的手段は、その情報に含まれるキーワードを抽出し、重み付けや統計的特徴を利用してユーザを分類することである。その際の分類指標として本研究ではいくつかの方法を提案しているが、本稿ではコサイン類似度※1を用いた相関値による性別推定について紹介する。

元となるデータとして、ツイート内容やフォロー関係等、2011年6月11日から2012年1月27日の期間収集したものを用いた。データの詳細内容を表1に示す。

判別のための規範データを作成するため、上記収集データセットから目視で性別の区別がつくユーザを男女各500名ずつ用意した。この用意にあたっては、ユーザの自己紹介分やツイートを目視し、性別を明らかに判断できる記述があるものに関して区別を行った。(表2参照)

推定の手がかりとなるキーワードを用意するために、まず始めに男女全ユーザ1,000人のツイートを形態素解析で素性に分解する。素性に分解した後、素性ごとに出現頻度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と出現回数を算出する。これらの値から、ベイズ推定と呼ばれる指標を利用して、あるキーワードが、男性がよく使うものか、女性が使うものかを算出する。算出した結果が、表3であり、「男性らしさ」の項目が1に近いほど、男性がよく使う言葉であり、0に近ければ、女性がよく使う言葉と解釈される。ここで、「男性らしさ」が0.6以上、または0.4以下の値をとり、かつ男女各々で出現する回数が多い素性を規範キーワードとしてピックアップした。また、表4に「男性らしさ」の範囲ごとにおけるキーワードの数について示す。表4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。

性別が不明なユーザに対して推定をおこなう手順としては、判別したいユーザのツイート等に含まれるキーワードと、表3で選んだ規範キーワードとの間で、コサイン距離に基づく類似度を算出して推定を行うものとした。

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	3,913
平均フォロー数	203
平均フォロワー数	235
延べツイート数	107,606,168
ジオタグ付きツイート数	279,101（全ツイートの0.26%）

表2　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

※1　コサイン類似度
ベクトル空間モデルにおいて、文書同士を比較する際に用いられる類似度計算手法。コサイン類似度は、そのまま、ベクトル同士の成す角度の近さを表現するため、三角関数の普通のコサインのとおり、1に近ければ類似しており、0に近ければ似ていないことになる。

このアルゴリズム※2の性能を測るために、適合率と再現率、そしてその調和平均であるF１指標※3と呼ばれる値を算出し評価した。

推定実験では、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規範キーワードの算出に利用し、残りの２つを組合せた200名について正しくプロファイルを推定できるかどうかの実験を行い、各クラスタの組み合わせによるクロスバリデーション※4によりモデルの汎用性を検証した。

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５のようになった。表５から、適合率は男性･女性とも80％を超えており、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、性別の推定が高い精度で行えることが明らかとなった。

5　まとめ

本研究では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を行い、本稿においては性別の推定結果を示した。

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、他のプロファイルの推定についても調査を行い、Twitterを利用するユーザ層がどのような方々なのか、その理解のもとに、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へと発展させ、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。

表３　性別を区別する際に用いた自己紹介文の例

キーワード	男性らしさ	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４　各重みの範囲におけるキーワード数と頻度

男性らしさ P の範囲	範囲に含まれるキーワード
0.9 ＞ P	2
0.9 ≦ P ＜ 0.8	6
0.8 ≦ P ＜ 0.7	17
0.7 ≦ P ＜ 0.6	71
0.4 ≦ P ＜ 0.3	435
0.3 ≦ P ＜ 0.2	109
0.2 ≦ P ＜ 0.1	71
0.1 ≦ P	42

表５　性別推定の結果

	適合率（%）		再現率（%）	
	男性	女性	男性	女性
クロスバリデーションの検証平均値	94.5	81.7	78.6	95.5
標準偏差	1.7	2.0	2.6	1.5
F 1 指標	男性：0.858		女性：0.881	

※２　アルゴリズム（algorithm）
ある特定の問題を解いたり、課題を解決したりするための計算手順や処理手順のこと。
※３　*F*1指標
適合率と再現率はトレードオフの関係にあるので、*F*1指標＝（2×適合率×再現率）／（適合率＋再現率）として、性能を測る。
※４　クロスバリデーション（Cross-validation）
交差検定。標本データを分割し、その一部をまず解析して、残る部分を最初の解析の仮説検定に用いる手法。