Twitter のフォロー関係のユーザの意図に基づく分類

田中　淳史†　　田島　敬史†

† 京都大学大学院情報学研究科社会情報学専攻 〒 606-8501 京都市左京区吉田本町
E-mail: †a.tanaka@dl.kuis.kyoto-u.ac.jp, ††tajima@i.kyoto-u.ac.jp

あらまし　Twitter はユーザによって利用目的が異なるため，あるユーザがあるユーザをフォローする際の意図も様々である．そこで本論文では，各フォローの意図をユーザ指向，内容指向，相互性の 3 つの軸で自動分類する手法を提案する．提案手法では，まず，ユーザを，その利用目的に基づいて大きく二つのタイプへと分類し，次に，各ユーザの個々のフォローの分類を，二つのユーザタイプ毎に個別に学習を行った SVM を用いて行う．さらに，SVM の属性として，各ユーザに関する情報に加え，Twitter のリスト機能において，各ユーザがコミュニティ型のリストと情報収集型のリストのどちらから多く参照されているか，また，情報収集型のリストの中でも，ユーザ指向的なリストや内容指向的なリストからどの程度フォローされているかの情報を用いる．これら二つの工夫により，単純に全ユーザのフォローを各ユーザの情報を属性として用いた一つの SVM で分類するのに比べて，精度の向上を実現した．

キーワード　マイクロブログ，Twitter

1. はじめに

近年，Twitter という Web サービスが注目を集めている．これはマイクロブログと呼ばれる新しいジャンルのサービスで，「ツイート」と呼ばれる 140 文字以内の短い文章を Web に投稿することで，ユーザが情報を発信したり，ユーザ同士でコミュニケーションを取ったりすることができるサービスである．Twitter は，今までのコミュニケーションサービスにはないフォローという新しい仕組みを持っている．

このフォローという仕組みにより，Twitter 上のユーザの繋がり方が様々なものになっている．その人にとって，友人であったり，有名人であったり，情報を発信するユーザであったり，様々なユーザがいるが，これらの様々なユーザを自由にフォローすることができる．そのため，ユーザがあるユーザをフォローする時の意図，すなわち，そのユーザへのフォローが何を目的としているのかは，そのフォローごとに異なり様々である．そして，フォローの意図はユーザ指向，内容指向，相互性という 3 つの軸で分類できるとわれわれは考えている．ユーザ指向とは，「そのユーザが発信するから内容を知りたい」というフォローの指向であり，内容指向とは，「発信するユーザによらず，そのような発信する内容が知りたい」という指向である．そして，相互性とは，「コミュニケーションを取るためのフォローであるかどうか」である．例えば，bot のような情報を投稿するユーザをフォローしている場合は，そのユーザの状態や考えを知りたいわけではなく，特定の分野の情報を知りたくてフォローしている場合が多いので，このフォローは内容指向の高いエッジであるといえる．また，友人をフォローしている場合は，コミュニケーション目的であり，かつそのユーザの状態や考えを知りたい場合が多いので，そのエッジはユーザ指向で相互性の高いエッジであるということがいえる．また，有名人であれば，そのエッジはユーザ指向が高く，かつ相互性が低いエッジであるということがいえる．本研究では，このようなユーザ指向，内容指向，相互性という三つの分類軸でフォローの意図を自動分類する手法を提案する．

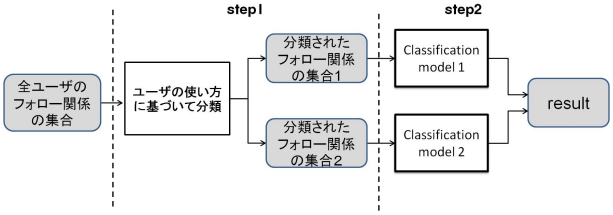

図 1　分類の手順

また，Twitter は様々な目的，利用形態で利用されている．例えば，報道のようなニュースについての投稿は，広い範囲に向けての情報発信を目的としており，blog のような使い方であるといえる．日記のようなユーザ自身の情報の報告は，コミュニティに発信することを目的とした，SNS のような使い方であるといえる．また，特定のトピックやテーマを意識して，メッセージを交換，収集することは，掲示板や RSS のような使い方であるといえる．さらに，ダイレクトメッセージ（フォローされている特定のユーザーに，第三者から見ることができないツイートを送る機能）を用いてのメッセージ交換は，IM のような使い方であるといえる．

このように，Twitter は様々な目的で利用可能なため，ユーザ毎の Twitter の利用目的も様々である．そして，利用目的が異なるユーザでは，フォローの意図の傾向が大きく異なる．そこで，本研究ではこの点に着目し，まずフォローする側のユーザを，主な利用目的がコミュニケーションであるユーザと，掲示板や RSS のように情報収集のために利用しているユーザに分類する．フォローする側のユーザの利用目的から，フォローの意図を二つの集合に分類し，そして，その二つのフォローの意図集合を，異なる学習器を用いて分類することで精度の向上を実現する．

また，本稿ではリストの分類を行い，その結果をフォローの意図の分類に用いることで，フォロー意図の分類のさらなる精

度向上を実現する．Twitter における全てのリストは情報収集型リストとコミュニティ型リストに分けることができる．情報収集型リストとは，ユーザがあるカテゴリやあるトピックの情報を収集するために，ユーザをまとめたリストである．例えば，鳩山由紀夫や舛添要一などの政治家を集めたリストや bot などをまとめたリストも情報収集型リストである．コミュニティ型リストとは，友人やその土地の人間を集めたリストなど，コミュニティの情報収集やコミュニケーションを行うためにユーザを集めたリストである．例えば，コミュニケーションを取るために，自分が所属するコミュニティ内のユーザや出会った人などをまとめたリストはコミュニティ型リストである．本稿では情報収集型リストとコミュニティ型リストに関する基礎的考察と，それらを分類する手法を提案し，実験評価を行う．

さらに，情報収集型リストをユーザ指向的リストと内容指向的リストに分類することを行う．リストにはユーザ指向で頻繁にフォローされるユーザが集まっているユーザ指向的リストと，内容指向で頻繁にフォローされるユーザが集まっている内容指向的リストがある．情報収集型リストはこの２つのどちらかの傾向を持っている．このリストの傾向の分類を行うために，提案手法ではリスト名を用いる．リスト名はユーザに対して付けられた，そのユーザの集まりを表現するアノテーションである．リストに付けられる名前によって，リストの中のユーザの集まりの傾向がユーザ指向的か内容指向的かが違うことが我々の調査から分かっている．本稿では，この情報収集型リストのユーザ指向度合いと内容指向度合いを求める手法を提案し，実験評価を行う．また，リスト分類結果の情報を，そのリストから参照されているユーザの情報として加えることで，フォローの分類の精度向上を実現する．

また，本研究で実現するフォローの意図の分類は，ユーザのリコメンデーションに応用できると考えられる．現在のユーザリコメンデーションシステムはユーザとユーザの興味の近さをもとに，リコメンデーションするユーザを決めている．しかし，われわれの分類結果を使えば，ユーザ間のトピックや関心の近さだけではなく，そのユーザがユーザ・内容・コミュニケーションのどの部分に興味を持ってフォローしているのか，そのユーザがフォローしている各ユーザは，その人にとって有名人であるのか，情報発信するユーザであるのか，友人であるのか，どのような人であるかという情報も利用して，リコメンデーションすることが可能になる．

本研究の寄与を以下に示す．

- 実際の Twitter データの分析から，フォローの意図やリストの使われ方にどのようなものがあるか示す．
- フォローの意図やリストの使われ方を自動分類する手法を示す．

また，本論文の構成は以下の通りである．第 2 章で関連研究について述べる．第 3 章では Twitter ユーザのタイプを分析し，そこから本稿ではどのようにフォローの意図を分類すべきと考えるかについて述べる．第 4 章でフォローの意図の分類に用いる属性について述べる．第 5 章では，リストの分類手法について提案し，実験評価を行う．第 6 章で実験の概要と評価について述べる．第 7 章でまとめと今後の課題について述べる．

2. 関 連 研 究

Twitter のデータを分析して，なんらかの情報を抽出・分類する研究がいくつかある．[1] は Twitter のネットワーク構造からコミュニティを抽出し，そこからユーザが twitter を使う目的について分析した研究である．この研究は，フォローの意図を分類するという本研究とは目的が大きく異なる．他には，そのユーザが人間であるか，サイボーグであるか，bot であるかを分類する研究 [2] スパムユーザを分類する研究 [3] などがある．これらの研究はユーザの特性から，フォローの種類を分類するという点で本研究と似ているが，本研究ではそのフォローの意図にもとづいて，より細かく分類することを目的としている．

さらに，Twitter がどのような使い方をされているかを分析している研究がいくつかある．[4] では，Twitter のユーザ数や使われている地域のユーザ数，投稿元などの統計データを詳細に集めている．また, [5] は，そのユーザがフォローされているリストの情報とそのリスト内のメンバーの情報から，そのユーザの特徴となる語を抽出している．

[6] は，ユーザの特徴とツイートの類似度から，ユーザのリコメンデーションを行う研究である．本研究のフォローの意図の分類結果を応用してリコメンデーションすることで，ネットワークや特徴の近さだけでなく，その人にとっての友達や興味のある情報を発信するユーザや興味のある有名人など，カテゴリに分けてリコメンデーションすることが可能になる．

3. ユーザの特徴と分類

この章では，Twitter 上のユーザのタイプとフォローの意図はどのような要素で構成されるかについて述べる．Twitter 上には様々なタイプのユーザがいる．有名人でありながら情報発信的なユーザもいれば，一般的ユーザであるが情報発信のみを行うユーザもいる．有名人のユーザの例として，鳩山由紀夫元首相 (hatoyamayukio) が挙げられる．鳩山由紀夫元首相のツイートは，特定のトピックに関する専門的な情報というよりは，一般的な意見や感想のみを述べたものが多く，鳩山由紀夫をフォローする人はそのユーザ自身に関心・興味があってフォローしている場合が多い．このように，発信内容に関係なくそのユーザによる発信であるという理由から多くのフォローを受けているユーザは，有名人である．

情報発信のユーザの例としては，bbcbusiness のようなユーザがいる．これらのユーザはニュース報道のような，事実や事象をツイートしている．短いコメントと共に，より詳細な情報のページの URL が書かれていることも多い．bbcbusiness などの大手メディアに限らず，個人のユーザの中にも情報発信指向のユーザはいる．これらのユーザは自分の詳しい分野に関しての情報を発信するメディアのような働きをするユーザである．また，Twitter におけるもっとも一般的なユーザとしては，コミュニケーションや情報収集を指向して Twitter を利用しているユーザが多く存在する．ツイートは日常的な会話などを発信することが多く，仲間内でのコミュニケーションが第一の目的であることが多い．スパマーとは，双方向フォローをされることを狙って，無差別にフォローするユーザである．エヴァンゲリストとは，自分の考えや情報を広めたいと思っているユー

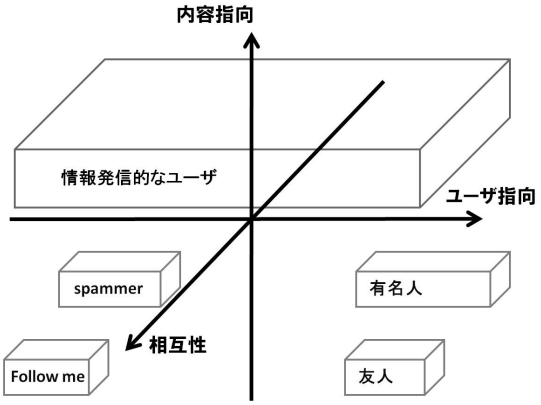

図 2　指向におけるフォローの意図

ザのことで，フォローされ返されることを狙ってフォローするユーザである．通常は，たくさんのユーザをフォローして，フォローされ返されることを待っている．これらのユーザの例としては，webdbf や maemukinews など，企業の広告的なユーザや情報を広めたいと思っている個人ユーザなどがいる．follow me と呼ばれているユーザもエヴァンゲリストの一種である [7]．このタイプのユーザは，フォロー数やフォロワー数を増やして，ネットワーク上のつながりが増えることを指向する．

Twitter 上には，このように多様なユーザが存在するため，ユーザがあるユーザをフォローする際の意図は様々なものがある．まず，Twitter がソーシャルネットワークサービスとしての側面も持つことから，従来のサービスと同じように，そのユーザとコミュニケーションを取りたいという意図のフォローが存在する．さらに，フォローという仕組みが RSS のような使い方もできることから，ネット上で発信されている情報を収集したいという意図のフォローが存在する．フォローの意図に関するこれらの調査・考察から，ユーザのフォローの意図の空間は，

- ユーザ指向
- 内容指向
- 相互性

の三つの次元によって構成されると考えられる．例えば，あるユーザが有名人である鳩山由紀夫をフォローしている場合，フォローしている理由はそのユーザの考えや思いを知りたいからであり，そのユーザ自身に興味・関心があるからである．友人をコミュニケーションのためにフォローしている場合も同じで，友人から実用的な情報を得ようというわけではなく，そのユーザ自身に興味・関心があるからフォローしている．このように，そのユーザであるから投稿する内容を知りたいという意図でフォローしている場合はユーザ指向のフォローであるといえる．そして，そのユーザ自身への興味でなく，そのユーザの発信内容を嗜好してフォローする場合は内容指向のフォローであるといえる．例えば，bot（定期的にニュースを投稿したり，遊びを目的として著名人や漫画・アニメのキャラクターを模倣した投稿を行う自動プログラム）のようなユーザをフォローしている場合は内容指向のフォローである．また，bbcbusinessnews のようなマスメディアへのフォローや，特定のトピックの情報が知りたくて，あるユーザをフォローしている場合も，内容指向のフォローであるといえる．ユーザ指向は「そのユーザが発信するから内容を知りたい」という指向であり，内容指向は「発信するユーザによらず，そのような発信する内容が知りたい」という指向である．両者の違いは，発信者に興味があるのか，発信される内容に興味があるのかの違いである．

そして，もう一つの軸として相互性がある．Twitter では，投稿の文頭に「@ユーザ名」を付けると，その人に宛てた投稿という意味になり，第三者ユーザーのタイムラインには表示されないが，双方をフォローしているユーザーのタイムラインには表示される．この「@」を用いた投稿により，互いの投稿内容にリプライすることができる．例えば，友人同士で，双方向に情報共有したりリプライしあったりしてコミュニケーションを楽しむことを目的としたフォローは，相互性のあるフォローといえる．以下では，相互性があるフォローとは，双方のユーザがコミュニケーションをとる間柄であり，コミュニケーションを取るためのフォローであると定義する．

図 2 は，あるタイプのユーザをフォローする場合のフォローの意図の例を，フォローの意図空間上に示したものである．友人であるからフォローする場合は，ユーザ指向であり，相互性があるが，内容指向ではない．有名人であるからというだけの理由でフォローする場合は，ユーザ指向であるが，内容指向と相互性はない．ユーザ指向でも内容指向でもないフォローには，スパマーやエヴァンゲリストなどによるフォローがある．これらのユーザは，自分が流したい情報を流すためだけに，たくさんの無差別なフォローを繰り返しているので，相手のユーザがどのような内容を発信していようと関係がなく，ユーザ指向でも内容指向でもない．また，スパマーはフォロー返しされて情報発信することを望んでいるが，コミュニケーションを目的としてフォローしているわけではないので相互性はない．しかし，エヴァンゲリストは情報交換や情報発信が目的であるため，エヴァンゲリストのフォローはコミュニケーション目的であり相互性があるといえる．情報発信的なユーザをフォローする場合，そのフォローは内容指向である．そして，情報発信的なユーザへのフォローがユーザ指向や相互性も持つかに関しては，その情報発信的なユーザの特徴による．例えば，bbcbusiness のようなユーザへのフォローであれば，そのユーザに興味があるわけではなく，さらに，コミュニケーションがとれるわけではないので，ユーザ指向も相互性もないフォローであるといえる．一方，内容指向のフォローでフォローされるユーザ，通常，情報発信的なユーザであるが，そのユーザの特徴によっては，ユーザ指向や相互性も持つフォローである場合もある．例えば，一般的なユーザが情報をよく投稿している場合，そのユーザをフォローする意図は内容指向であるが，そのユーザと知り合いであれば，コミュニケーション指向のフォローともいえるので相互性がある．また，そのユーザと知り合いであるので，ユーザ指向のフォローであるともいえる．このように情報発信的なユーザをフォローする場合は，内容指向であるということは共通するが，ユーザ指向があるか，相互性があるかは，フォローするユーザの特徴に依存する．

フォローの意図の分類軸としては，これら 3 つ以外にも有用なものが存在する可能性はあるが，上で示したように，これらの 3 つの分類軸を用いると，Twitter 上で見られるフォローの意図のうちの主要なものを効果的に分類することができる．そ

こで，本稿では，フォローの意図をこれらの 3 つの分類軸に従って分類することを考える．他の有用な分類軸としてはどのようなものが考えられるかの検討，また，そのような軸に関する自動分類手法の設計については今後の課題である．

4. フォローの分類に用いる属性

この章では，フォローの意図を，SVM を用いて前述の三つの軸で自動分類する際に，重要であると考えられる属性について説明する．表 3 は，本研究で重要と考える属性をまとめたものである．ここでは，ユーザ A がフォローされる側でユーザ B がフォローする側である．B から A へのフォローの意図を推定する場合，ユーザ A がツイートの投稿に関してどのような特徴を持っているユーザであるかと，ユーザ B がどのようなフォローをしているユーザであるかが重要となる．さらに，ユーザ A とユーザ B の間にどのような関係があるかも重要である．そこで，ユーザ A とユーザ B の各々の属性の他に，相互関係を表す属性も加えている．

以下，これらの属性について順に説明していく．まず，基本的な属性としてユーザ A とユーザ B のフォロー数，フォロワー数，フォローされているリスト数の情報を用いる．さらに双方向フォロー数と双方向率を属性として加える．双方向フォロー数とは，あるユーザが相互にフォローしあっている相手の数である．また，双方向率は，そのユーザがどの程度コミュニケーション指向のユーザであるのかを推定する上で重要な属性であり，ここでは，以下の式で定義される $mutual\_followee\_ratio$ と $mutual\_follower\_ratio$ の二種類の双方向率を用いる．また，そのユーザのフォロー数が 0 の場合は，$mutual\_followee\_ratio$ を 0 にする．

$$mutual\_followee\_ratio = \frac{\text{双方向エッジ数}}{\text{そのユーザのフォロー数}}$$

$$mutual\_follower\_ratio = \frac{\text{双方向エッジ数}}{\text{そのユーザのフォロワー数}}$$

二つの双方向率が共に高い場合は，そのユーザがコミュニケーション指向であり，そのフォロー先もフォロワーも友人ばかりである可能性が高い．$mutual\_followee\_ratio$ が低く $mutual\_follower\_ratio$ のみが高い場合は，有名人や情報発信をしているユーザをフォローしつつ，自分自身はあまり情報発信をしていないので，そのユーザのフォロー数が少ない場合はフォロワーは友人ばかりというユーザである可能性が高いが，フォロー数がある程度多い場合はフォロー返しをしてもらうことを狙うエヴァンゲリストである可能性もある．一方，$mutual\_follower\_ratio$ が低く $mutual\_followee\_ratio$ のみが高く，ある程度フォロワー数が多い場合は，一方向のフォロワーが多いという面から，有名人やなんらかの情報発信を行っているユーザである可能性が高く，また，フォローは主にコミュニケーションのためのみに行っていると考えられることから，コミュニケーションのためにも Twitter を利用しているような有名人である可能性が高い．一方，$mutual\_follower\_ratio$ と $mutual\_followee\_ratio$ がともに低い場合は，たくさんのフォローをして情報収集をしつつ，なんらかの情報発信をしていてフォロワーもたくさんいるような一般の個人の情報発信型ユーザである可能性が高い．

共通リスト数は，ユーザ A とユーザ B を共にフォローしているリストの数のことである．ユーザ A と B に関して，そのようなリストの数が多い場合は，これらのユーザが同一のトピックに関する有名な情報発信ユーザや bot である可能性がある．また，ほぼ必ずもう一方もフォローしているというような，共通リスト率が高い A と B の組み合わせは，同じコミュニティに属している知り合いであることが考えられ，相互性がある可能性が高い．また，知り合い同士の間のフォローは，ユーザ指向であることが多い．そして，有名人や情報発信のユーザと同じリストからフォローされているユーザは，そのユーザも有名人や情報発信のユーザである可能性が高い．これらのことから，各々をフォローしているリストの数と，両者を共にフォローしているリストの数が，ユーザ同士の特徴の近さを知る上で有用な情報となることが分かる．

「双方向フォローであるか」という属性の値はお互いにフォローしあっている場合を 1 とし，そうでない場合を 0 とする．分類軸の相互性の定義は，コミュニケーションがとれる間柄であるかどうかなので，相互フォローの関係であっても，bot であったり，コミュニケーションを目的としてフォローしていない場合は相互性がないといえる．つまりこの属性は，相互性があるための必要条件であるが十分条件ではない．また，@マークによる返信の頻度は，高いほど相互性が高いといえる．また，そのユーザと知り合いである可能性が高いことから，ユーザ指向である可能性が高いということも言える．ユーザ A をユーザ B が RT（他のユーザのツイートを自分のフォロワーに向けて再投稿すること）する頻度は，内容指向に影響があると考えられる．

フォローの意図の傾向（ユーザ指向，内容指向，相互性）は，一旦，この情報を用いずにフォローの意図を分類し，その結果において，あるユーザがどのような傾向で他のユーザをフォローしていることが多いかの傾向を表す属性である．例えば，あるユーザ u が n 人のユーザ $f_1, \ldots, f_n$ をフォローしていて，ユーザ f_i へのフォローの意図の，各軸に関する判定結果（区間 $[0,1]$ 内の値で与えられる）をそれぞれ $u(f_i), t(f_i), m(f_i)$ とすると，u のフォローの各軸に関する傾向は，以下のようにその軸に関する評価値の平均で定義される．

$$Tend_u(u) = \frac{1}{n} \sum_{1 \leqq i \leqq n} u(f_i)$$

$$Tend_t(u) = \frac{1}{n} \sum_{1 \leqq i \leqq n} t(f_i)$$

$$Tend_m(u) = \frac{1}{n} \sum_{1 \leqq i \leqq n} m(f_i)$$

ここでは，ユーザ B の，$Tend_u, Tend_t, Tend_m$ を計算し，属性に加える．この属性によって，ユーザごとの Twitter の利用目的の違いの情報を，SVM における学習の段階でも利用することができると考えられる．

表 3 にあげられた属性のうち，ここで説明していないリストに関する属性については，5 章のリストの分類の中で説明をする．これらの属性による分類の実験，評価は 6 章で示す．

5. リストの分類

この章では，Twitter のリスト機能の概要を説明し，本研究で提案するリストの分類手法を述べる．また，リストを分類した結果がフォローの意図のの分類にどのように寄与するかについても述べる．

5.1 情報収集型リストとコミュニティ型リスト

Twitter では，フォロー数が多くなると，自身のタイムラインの情報の流れが速くなり，閲覧したい情報を逃すことが多くなる．そのような場合に，フォローしているユーザをグループわけして，グループごとにツイートを表示する機能がリスト機能である．そして，各リストには作成者によって名前が付けられるため，リスト名がそのリストのメンバーの特徴や興味を反映していることが多い．

例えば，politics という名前のリストには，鳩山由紀夫などの実際の政治家がメンバーとして登録されていることが多く，また，news という名前のリストには，ニュースをツイートする bot ユーザが登録されていることが多い．このような，ある分野のオーソリティであるような有名人ばかりを集めたリストもあれば，有名人ではないが，あるテーマに関する情報発信をしている一般のユーザを集めたリストもある．例えば，Ruby という名前のリストには，Ruby に関するオーソリティ的なユーザのみならず，頻繁に Ruby に関する情報を発信する一般の Ruby 開発者もメンバーとして登録されていることがある．これらは，情報収集型のリストである．

一方, conversationlist などの名前のリストや, kyodai_jouhou のような，そのユーザが属している組織，グループの名前が付けられているリストは，そのユーザが Twitter 上でコミュニケーションを取るユーザを集めたものであることが多い．これらは，コミュニティ型のリストである．また，先ほど例に挙げた「Ruby」などの名前のリストであっても，そのユーザ自身が Ruby に関する情報発信をしているユーザであり，そのユーザが日常的に Ruby に関する議論や相互の情報共有をしている知り合い，あるいは常連のユーザを集めたようなリストであれば，コミュニティ型リストの側面も持っていると言える．

このように，リストにはある情報を収集するためにグループ分けしたのものと，コミュニケーションやコミュニティ間の情報共有を目的としてグループ分けしたコミュニティ型の 2 種類に分けられる．

5.2 情報収集型リストとコミュニティ型リストの分類

この節では情報収集型リストとコミュニティ型リストを分類する手法を述べる．Twitter 全体の 80%のユーザがフォロワー数よりもフォロー数の方が多く，有名人やある分野のオーソリティなどのように，フォロー数よりフォロワー数が多いのは全体の 20%しかいないことが分かっている [8]．そのため，情報収集型リストのメンバーのフォロワー数の合計は同じメンバー数のコミュニティ型リストのフォロワー数の合計数より多い傾向にあることがわれわれの調査で分かっている．そこで，情報収集型リストであるかコミュニティ型リストであるかを特定するために，各リスト l について以下の式で定義される値を用いる．

$$listrate(l) = \frac{1}{|member(l)|} \sum_{u \in l} \frac{followed_u}{following_u}$$

表 1　情報収集型リストとコミュニティ型リストの $listrate$

$listrate(l)$	コミュニティ	情報収集型
平均	25.3	2694.8
標準偏差	79.5	5057.1
中央値	0	571
最小	0	0
最大	471	21534
信頼区間 (99%)	30.5	1959.5

表 2　リスト分類の結果

	情報収集	コミュニティ
適合率	84.6%	88.8%
再現率	89.7%	83.3%

ここで，$member(l)$ は l のメンバー集合，$following_u$ はメンバー u のフォロー数，$followed_u$ はメンバー u のフォロワー数である．この式で，どの程度リストメンバーの中にオーソリティや有名人のようなユーザがいるのかがわかる．また，オーソリティや有名人をフォロワー数から発見することが目的であるので，$\frac{followed_u}{following_u} < 1$ のユーザの値は 0 として，計算を行う．また，$following_u$ が 0 の場合は計算できないので，値を $following_u = 1$ として計算を行う．この式で計算を行うと，情報収集型のリストであれば，$listrate$ は高くなり，一般ユーザばかりいるコミュニティ型のリストであれば，$listrate$ は低くなる．そこで，以下の条件を用いて，リスト l が情報収集型リストであるかコミュニティ型リストであるかの判定を行う．

$$\begin{cases} \text{if } listrate(l) > \Theta & l \text{ は情報収集型リスト} \\ \text{otherwise} & l \text{ はコミュニティ型リスト} \end{cases} \quad (1)$$

先ほどクロールしたリストから，ランダムにそれぞれコミュニティ型リストと情報収集型リストを 100 件ずつ選び，これらの $listrate$ を計算した結果が表 1 である．コミュニティ型リストにおいては $listrate$ の中央値が 0 であることから，ほとんどのリストのメンバーがフォロワー数よりフォロー数が多いことが分かる．また，コミュニティ型リストの $listrate$ の最大値は 471 であった．情報収集型リストの値の平均を見てみると，2694.8 とコミュニティ型リストより高くなっていることがわかる．中央値も 571 とコミュニティ型リストの最大値より高い．

上の結果から，$listrate$ を用いることで，コミュニティ型リストと情報収集型リストをある程度わけることができると考え，$listrate$ を用いた分類の実験を行った．我々は実験結果から，Θ は 4 が最適であると考えた．このことから，$listrate$ の Θ が 4 を超えるものを情報収集型リスト，4 以下のリストをコミュニティ型リストとして分類する．収集したリストの中からランダムで 100 件のリストを選び，それぞれ $listrate$ を計算して，その値から情報収集型リストとコミュニティ型リストに分類し，検証した結果が表 2 である．

情報収集型リストの分類では，適合率は高いが再現率が低い．逆に，コミュニティ型リストの分類では，再現率は高いが適合率は低いという結果になっている．しかし，適合率がトピック・コミュニティ共に 80%を超えているので，$listrate$ を計算するという簡単な手法で，ある程度，情報収集型リストであるかコミュニティ型リストであるかを分類できるといえる．

そして，この分類結果を，SVM によるフォローの分類の際の，各ユーザの属性として加える．表 3 の情報収集型リストとコミュニティ型リストのフォロー比がそれである．本研究の実験では，ユーザがフォローされているリストから最大 20 件をランダムに選び，それらの中の情報収集型リストとコミュニティ型リストの比をユーザの属性として加えている．

5.3 ユーザ指向的リストと内容指向的リストの分類

この節では，リスト名の情報を用いて，情報収集型リストを，さらに，ユーザ指向的なものと内容指向的なものに分類する手法について述べる．リスト名はそのメンバーに関連するトピックや単語による名前が付けられていることが多い．例えば，bot などの名前がついたリストであれば，リストの中のユーザの多くは bot であり，そのリストは内容指向的である可能性が高い．一方，famous のような名前のリストであれば，リストの中のユーザの多くは有名人であり，ユーザ指向的なリストである可能性が高い．このように，リスト名を用いることで，そのリストがユーザ指向的か内容指向的かを判定することが可能だと考えられる．

しかし，そのような特徴的なリスト名の数は限られている．そこで，本研究では，さらに，4 章でのフォローの意図分類の結果をこの分類にも利用することを考える．まず，われわれの調査では，ユーザ間のフォローの意図は，フォローされる側のユーザの性質に，より依存する傾向がある．そこで，あるユーザへのフォロー群が，ユーザ指向か内容指向かの推定結果（リストの情報を用いずに，他の属性のみを用いて，一旦，推定した結果）を表すベクトル値を，そのユーザがユーザ指向でフォローされることが多いユーザか内容指向でフォローされることが多いユーザかの傾向を表す属性値として，そのユーザに与える．そして，リストとユーザの間のフォロー関係の二部グラフを利用して，これらの値を伝播させ，リスト名ごとのユーザ指向と内容指向の度合いを求める．

具体的な手順の概要は以下のようになる．

(1) 各ユーザへのフォローの意図分類の結果を表すベクトル値をそのユーザに与える
(2) 各ユーザをフォローしているリストを収集する
(3) 同じ名前を持つリストは一つのリストにマージする
(4) 各ユーザの持つ値を，そのユーザをフォローしているリストに伝播させる
(5) 各リストの値を，そのリストがフォローしているユーザに伝播させる
(6) 上の (4), (5) を繰り返す

手順 (1) では，あるユーザ u へのフォローの意図の推定結果を，$U_U(u) = (u(f_1), u(f_2), \ldots)$，$C_U(u) = (t(f_1), c(f_2), \ldots)$ という二つのベクトルの形で表し，これを u の属性の初期値として与える．ここで，$f_1, f_2, \ldots$ は，u のフォロワー，$u(f_i)$ はフォロワー f_i の u へのフォローがユーザ指向であるかを表す $0, 1$ の値，$t(f_i)$ はフォロワー f_i の u へのフォローが内容指向であるかを表す $0, 1$ の値である．

手順 (4) では，各ユーザが持つ上記の値を（リスト名毎にマージされた）各リストに伝播させるために，まず，リスト l のユーザ指向度と内容指向度の値を表すベクトル $U_L(l)$ と $C_L(l)$ を以下の式で求める．

$$U_L(l) = \frac{1}{|member(l)|} \sum_{u \in member(l)} U_U(u)$$

$$C_L(l) = \frac{1}{|member(l)|} \sum_{u \in member(l)} C_U(u)$$

さらに，ここでは，リストは必ずユーザ指向的なものか内容指向的なものかどちらか一方であると仮定し，以下の式により，$U_L(l), C_L(l)$ のうち値が小さい方を 0 とする．

$$\begin{cases} U_L(l) > C_L(l) & C_L(l) = 0 \\ \text{otherwise} & U_L(l) = 0 \end{cases}$$

同様に，手順 (5) ではリストの持つ値をユーザへ伝播させるために，まず，以下の式により $U_U(u), C_U(u)$ を更新する．

$$U_U(u) = \frac{1}{|list(u)|} \sum_{l \in list(u)} U_L(l)$$

$$C_U(u) = \frac{1}{|list(u)|} \sum_{l \in list(u)} C_L(l)$$

ただし，ここで，$list(u)$ はユーザ u をメンバーに持つリストの集合である．また，ユーザについてはユーザ指向でフォローされる傾向，内容指向でフォローされる傾向の双方を持ちうると考えられるので，どちらか一方を 0 にすることはしない．

以上の (4), (5) を繰り返し行いたのちの収束値が各ユーザのユーザ指向的リストの度合いと内容指向的リストの度合いとなり，これらを，表 3 のユーザ A の側の属性として加える．

6. 実　　験

この章では，本研究で提案する，フォローの意図をユーザ指向，トピック指向，相互性の軸で自動分類する手法と，その実験の評価・考察について述べる．

6.1 実験データ

実験に用いるデータとして，19 人のユーザにそれぞれがフォローしているユーザをランダムに 40〜50 人選んでもらって合計 826 件のフォローのデータを集め，それぞれのフォローをユーザ指向，トピック指向，相互性の 3 つの軸に関して，当てはまる場合は 1 を，当てはまらない場合は-1 として評価してもらい，それぞれのクラスの学習データとした．また，なるべく異なった使い方をしているユーザを集めるために，19 人のユーザは以下のような方針で集めた．まず，幅広い使い方のユーザを集めるために，19 人のユーザのフォロー数は数十人のユーザ 8 人，数百人のユーザ 6 人，数千人のユーザ 5 人とした．さらに，ユーザ自身が知り合いのコミュニティ内でのコミュニケーションを指向しているのか，掲示板や RSS のような使い方をして広い範囲での情報発信・収集をしているのかを申告させ，前者のユーザを 11 人，後者のユーザを 8 人ずつそれぞれ集めた．フォローの意図分類の前段階であるユーザの分類に関しては，このユーザの申告結果を利用した．また，使い方や嗜好の偏りを防ぐために実世界上で異なるコミュニティに所属している人物を集めた．表**??**は評価データのユーザ数，ユーザ指向，内容指向，相互性の総数を示したものである。掲示板や RSS のような使い方をして広い範囲での情報発信・収集をしているグループを G1，知り合いのコミュニティ内でのコミュニケーションを指向しているグループを G2 としている．

6.2 実験方法

本実験では，教師あり学習を用いる識別手法の一つである SVM を使用して，自動分類を行う．本研究で使用した SVM ツールは LIBSVMversion2.91 であり，カーネルはガウスカーネルを使用した．10fold-crossvalidation で実験を行った．本実験で SVM に与える属性の一覧を表 3 に示す．なお，SVM の学習データとして与える属性値の範囲は [0,1] が適切なため，各属性の属性値は [0,1] の範囲に正規化して用いた．また，フォロー数とフォロワー数に関しては，膨大な値を持つユーザがいるため，対数をとった上で [0,1] に正規化した．これは，例えば，フォロー数 10 人と 60 人では大きくそのユーザの使い方が違うことが考えられるが，フォロー数が数万であるようなユーザがいると，この値の違いが反映されづらくなるので，このような極端に大きな値の影響を小さくするためである．

また，前述のように本研究の提案手法では，フォローの分類の精度を向上するために，まずユーザを，Twitter の利用目的に応じて分類し，各ユーザタイプ毎に異なる SVM を用いる．本実験では，まず，ユーザを，

- G1: Twitter を掲示板や RSS のように使って，興味のあるトピックや内容の情報を収集するために使ったグループ
- G2: コミュニケーションが Twitter の主な利用目的であるグループ

の二つに分けて，それぞれについて SVM による学習を行う実験を行い，これらの結果を合わせたもの（以下，G1+G2 で表す）と，そのような分類を行わずに一つの SVM で全フォロー関係の学習，分類を行った場合（以下，これを A1 で表す）の結果との比較を行う．また，表 3 に示した属性のうちの，さまざまな部分集合のみを用いた場合との比較評価も行う．表 5 に，比較実験を行った属性集合の一覧を示す．

6.3 実験結果と考察

図 3，4，5 は，それぞれ，ユーザ指向，内容指向，相互性に関する提案手法による分類結果の精度をグラフにしたものである．精度は LIBSVMversion2.91 の出力するものを用いた．コミュニケーション指向である G2 のグループでは，runID A，すなわち，属性 (i) のみを用いるもっとも単純な手法でも，比較的，高い精度で分類できている．それに対して G1 のグループでは，A のみでは，精度がどの軸についても 70%前後と低い．また，G2 のグループは用いる属性を A〜K と増やしていっても，精度の上昇はゆるやかで大きく上がらないことがわかる．それに対して，G1 は I でリストの分類結果を加えたところで大きく上昇していることがわかる．

ユーザ A がユーザ B を RT する頻度と@マークの頻度は内容指向の分類にわずかに影響がある場合があるが，ほとんど精度に影響はないと考えられる．また，フォローの意図の傾向も精度には影響がないと考えられる．mutual_follower_ratio と mutual_followee_ratio に関しては G1 のグループでは，精度向上に繋がっている．G2 のグループは，コミュニケーション指向であり，有名人や情報発信的なユーザをフォローするにしても，ユーザによってフォローの意図の偏りが少ないと考えられる．そのため，フォローされる側の A の段階である程度高い精度が出ていると考えられる．また，両者のグループとも双方向エッジの属性があまり精度に影響がないことがわかる．このことから，エッジが双方向であるからといって，それがコミュニケーションを取りあうためのものではないということがわかる．

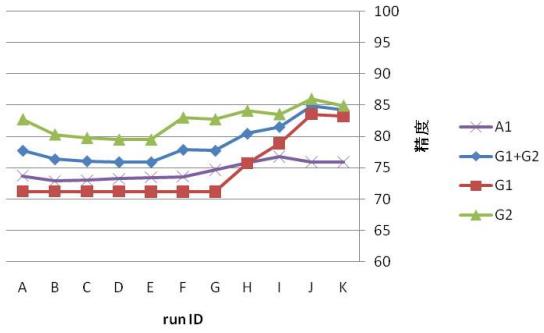

図 3　ユーザ指向の結果

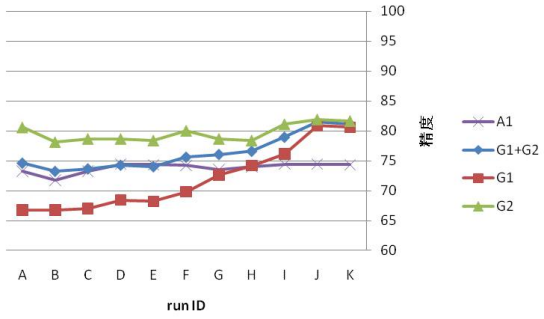

図 4　内容指向の結果

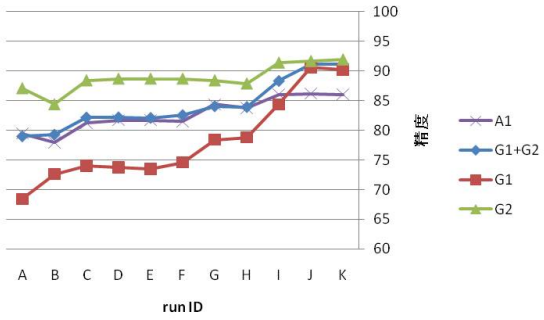

図 5　相互性の結果

また，Twitter の利用目的によってデータを分けた場合 (G1+G2) と分けなかった場合 (A1) では，分けた場合 (G1+G2) の方が大きく精度が向上していることがわかる．また，A1 では，I, J の段階でリスト分類の結果を適用しても精度向上に繋がらなかったのに対して，G1 と G2 のグループでは向上していることがわかる．これは，ユーザの使い方によって，リスト名の付け方の傾向が異なり，同じような使い方のユーザ間では，リスト名ごとのユーザ指向度合いと内容指向度合いが似ているためではないかと考えられる．

表 5　実験データの内訳

	ユーザ数	ユーザ指向	内容指向	相互性
全ユーザ	872	644	302	453
G1	379	274	142	185
G2	493	370	160	268

表 3　エッジ分類に用いる属性

	属性番号	属性
ユーザ A	(i)	フォロー数，フォロワー，フォローされてるリスト数，双方向エッジ数
	(ii)	mutual_follower_ratio，mutual_followee_ratio
	(iii)	情報収集型リストとコミュニティ型リストのフォロー比
	(iv)	ユーザ指向的リストからフォローされている度合い
	(v)	内容指向的リストからフォローされている度合い
ユーザ B	(vi)	フォロー数，フォロワー数，フォローされてるリスト数，双方向エッジ数
	(vii)	mutual_follower_ratio，mutual_followee_ratio
	(viii)	情報収集型リストとコミュニティ型リストのフォロー比
	(ix)	フォローの意図の傾向（ユーザ指向，内容指向，相互性）
相互	(x)	共通リスト数
	(xi)	双方向エッジであるか
	(xii)	マークによる返信の頻度
	(xiii)	ユーザ A をユーザ B が RT する頻度

表 4　属性と runID の対応

run	features												
	followee					follower				relation			
ID	(i)	(ii)	(iii)	(iv)	(v)	(vi)	(vii)	(viii)	(ix)	(x)	(xi)	(xii)	(xiii)
A	✓												
B	✓					✓							
C	✓					✓				✓			✓
D	✓					✓				✓		✓	
E	✓					✓				✓			
F	✓	✓				✓				✓			
G	✓	✓				✓	✓			✓			
H	✓	✓				✓	✓			✓	✓		
I	✓	✓	✓			✓	✓	✓		✓	✓		
J	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓		
K	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		

7. ま と め

本稿では，各ユーザがユーザをフォローする際に様々な意図があることを述べた．そして，それらのフォローの意図はユーザ指向，トピック指向，相互性の 3 つの軸で分類できると考え，SVM を用いてユーザ間のフォローの意図を分類する手法を示した．また，Twitter のリストには情報収集型リストとコミュニティ型リストがあることを示し，さらに，情報収集型リストの中には，ユーザ指向的リストと内容指向的リストがあることを述べた．そして，これらのリストの分類手法について示した．また，ユーザを Twitter の利用目的に応じて分類してからフォローの意図の自動分類を行うことや，リストの分類結果をフォローの意図の分類の際の属性として用いることで，分類精度が向上することを示した．

今後の課題としては，ユーザの分類を自動にすることである．今回の実験では，ユーザの申告をもとに分類を行ったが，ユーザの分類に適切な属性を与え，自動分類を行う必要がある．また，3 つの軸の他に有用な分類軸としてはどのようなものが考えられるかの検討，また，そのような軸に関する自動分類手法の設計についても今後の課題である．

文　献

[1] “Why we twitter: understanding microblogging usage and communities”, International Conference on Knowledge Discovery and Data Mining, pp. 56–65 (2007).

[2] “Who is tweeting on twitter: human, bot, or cyborg?”, Proceedings of the 26th Annual Computer Security Applications Conference, ACSAC ’10, New York, NY, USA, ACM, pp. 21–30 (2010).

[3] “Uncovering social spammers: social honeypots + machine learning”, Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval, SIGIR ’10, New York, NY, USA, ACM, pp. 435–442 (2010).

[4] “A few chirps about twitter”, Proceedings of the first workshop on Online social networks, WOSP ’08, New York, NY, USA, ACM, pp. 19–24 (2008).

[5] “Analysis of twitter lists as a potential source for discovering latent characteristics of users”, Workshop on Microblogging at the ACM Conference on Human Factors in Computer Systems. (CHI 2010) (2010).

[6] “Recommending twitter users to follow using content and collaborative filtering approaches”, Proceedings of the fourth ACM conference on Recommender systems, RecSys ’10, New York, NY, USA, ACM, pp. 199–206 (2010).

[7] “’follow me’: a web-based, location-sharing architecture for large, indoor environments”, Proceedings of the 19th international conference on World wide web, WWW ’10, New York, NY, USA, ACM, pp. 1375–1378 (2010).

[8] “What is twitter, a social network or a news media?”, Proceedings of the 19th international conference on World wide web, WWW ’10, New York, NY, USA, ACM, pp. 591–600 (2010).

[9] “Mining the link structure of the world wide web”, IEEE Computer (1999).

[10] “Spatio-temporal-thematic analysis of citizen sensor data: Challenges and experiences”, Proceedings of the 10th International Conference on Web Information Systems Engineering, WISE ’09, Berlin, Heidelberg, Springer-Verlag, pp. 539–553 (2009).