

twitter のツイートに関する分類手法の提案

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 は掲示板のような使い方, チャットのような使い方, blog のような使い方など様々な形態で利用されている。そのため, twitter に投稿されるツイートには, そのユーザの友人間などの特定のコミュニティ内でのみ意味があるものから, 広く一般の人にとって有用な情報までさまざまである。そこで, 本論文では, これらのツイートを, 広く一般の人にとって情報として有用なものとそうでないものに分類する手法を提案する。本論文で提案する手法では, 各ツイートが広く有用なツイートであるかどうかの判定を, そのツイートを投稿したユーザが, 一般への情報発信を指向しているユーザか, 主にコミュニティ内のコミュニケーションのために twitter を使用しているユーザかという点と, そのツイートの内容が一般への情報発信を指向しているかどうかという点の二つの観点に基づいて行う。また, 各ユーザが情報発信を指向しているかどうかの判定の手法として, そのユーザへのフォロー関係を, 特定のコミュニティ内のフォロー関係とそうでないものに分類することによって行う手法と, そのユーザのツイートが「お気に入り」に登録される頻度をを用いる手法を提案し比較する。

キーワード マイクロブログ

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。ツイートは Web 上に時系列で次々と発信することができ, 更新が容易で短い文章を投稿するため, blog や SNS よりも比較的手軽に文章を投稿できるという特徴がある。また, 気軽に投稿できることから, チャットや IM のようなリアルタイム性の高い内容の投稿がよくされる。twitter では, 掲示板や IM と似たインターフェースでツイートを閲覧するが, これまでのサービスとは閲覧の方式に違いがある。従来の掲示板やチャットのように「特定のテーマについての場に個人が集まる」という方式や, IM のように「メッセージを交換しあう」という方式ではなく, あくまで「個人の発信がそれぞれのページにオープンな状態で載っており, それを収集する」という閲覧方式である。この twitter の閲覧方式とサービスの特徴は, 今までのコミュニケーションサービスにないフォローという新しい仕組みによってもたらされている。フォローとは自分のホームページ上に他のユーザのツイートを表示させるための閲覧登録である。フォローすると, そのユーザのツイートが自動的に収集され, 自分のホームページで閲覧することが可能になる。ユーザをフォローしている人のことをフォロワーと呼ぶ。twitter ではフォローの許可を相手に取る必要がなく, 誰でもフォロワーになることができる。フォローという閲覧登録の形態は, 相手に許可を取らない点や, 必ずしも双方向に情報を発信・受信しあわない点が SNS と違う点である。また, フォローしなくても閲覧するだけならば自由に他のユーザのツイートを見ることができ

る^(注1)。他のユーザのツイートを閲覧して, それをもとにフォローするのも, しあうのも自由である。この自由な情報閲覧の形態により, 手軽な情報発信が実現されている。

現在, twitter 上のツイートはコミュニティ内でのみ意味のあるものから, 広く一般に有用なもので発信範囲が様々である。内容に関しても, リアルタイム性の高いツイートもあれば, 低いツイートもあるし, 何の脈絡もない突然のツイートもあれば, ディスカッションのようなツイートもある。このように, twitter 上で発信されるツイートの内容は様々である。これは twitter が, 従来の多くのコミュニケーションサービスと同じ使い方ができることで, そのようなサービスと同じ傾向の投稿が生まれるためであると考えられる。例えば, 報道のようなニュースについての投稿は, 広い範囲に向けて情報を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ており, 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, メッセージ交換をすることは, 掲示板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^(注2)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, 従来のサービスと同じような使い方ができることから, 情報発信・収集, 情報交換や, コミュニケーションなど様々な投稿が twitter 上で行われ, ツイートの内容は多様なものとなる。

また, この多様な内容のツイート中には有用な情報もあり, 最近では twitter で情報を収集したいというユーザの要求が高まっている。しかし, twitter が提供しているサービスのみで

(注1): ツイートをプライベートにしている, 登録の許可がないとツイートを開覧できないユーザもいる。

(注2): フォローされている特定のユーザーに, 第三者から見るできないツイートを送る機能。

は、有用なツイートを発見することは困難であるといえる。その理由は twitter の提供する情報収集機能がフォローとリアルタイム検索^(注3)のみであり、その2種類の方法が情報収集する上で非力であるからである。前者を行う場合には、自分のフォローやフォロワーからユーザをたどって^(注4)、それらのユーザの最近のツイートをざっと見て、自分にとって有益なユーザを発見し、さらにそのユーザのツイートを一つ一つ確認して、その中から有用な情報を見出す必要がある。twitter のユーザ数は膨大で、この方法で有益なユーザを発見することは手間のかかることである。ユーザにもよるが、twitter 上でのツイートのほとんどがユーザの個人的な情報である場合が多く、フォロワー以外のユーザにとっての有用な情報がツイートされる確率は低い。また、リアルタイム検索でも、そのトピックの単語を含むツイートをグローバルに集めてきて表示するので、自分の知りたい情報を含むツイートは不要なツイートの中に埋もれてしまうということが問題としてある。そこで、本研究では twitter 上のツイートの中から有用な情報のみを取り出すことを目的とするツイートの分類手法を提案する。

本研究では、ツイートの分類を行う前に、ユーザの分類を行う。ユーザのツイートの発信傾向は、それぞれのユーザの使い方の違いに大きく関係するからである。そして、ツイートだけで分類を行うことは、一つのツイートが持つ情報量が少ないことから困難であると考えられる。なぜなら、twitter の投稿内容は 140 文字と制限されており、ブログや Web ページと同じように、従来の文脈ベース解析を行うことはできない。ツイートの情報のみから正しく分類を行うことが難しいため、ユーザ分類を前段階として行う。ユーザ分類後の適当なユーザに対して、ツイート分類を行うことで、分類の精度を向上させることが狙いである。

ユーザ分類に関して、今回は2種類の提案手法で実験を行う。1つはフォロー関係の意味を考慮することで有用なユーザを判定する手法で、もう1つがお気に入りをつけられたツイートの割合から有用なユーザを判定する手法である。お気に入りとは、気に入ったツイートを登録することのできる機能である。以下では、2つのユーザ分類に関する提案手法で用いる、フォロー関係の意味とお気に入りについて説明する。

twitter 上でのリンク関係には意味がある。例えば、単純に友達だからフォローしている場合や、趣味のあうツイートをしてくれるからフォローしているという場合や、有名人であるからフォローしている場合など様々な理由が考えられる。それぞれのユーザのフォロー関係に意味が存在する。ツイートの分類を行う上で、我々が着目するのは有用な情報をツイートするユーザである。よって、ユーザ分類の時点で、そのようなツイートをするユーザを発見しなければならない。そのため、単に Web ページと同じように HITS アルゴリズムや PageRank を用いて、ネットワーク解析を行うよりも、そのユーザとフォロワー

のエッジの意味関係に着目して、ネットワーク解析を行ったほうが、より分類の目的に即したユーザを発見することが可能である。本研究では単にエッジとエッジの始点になっているノードの数の関係に着目して解析を行うのではなく、ユーザ同士がどのような意味を持ってフォローしているのかの関係を発見し、それを考慮した上でユーザを評価する。例えば、ユーザ同士がそれぞれをフォローしあっている場合、その関係は友人・知人関係である可能性が高いといえる。我々が分類において注目するユーザは、他のユーザへの情報発信意識が高く、有用な情報をツイートする割合が高いユーザである。そういったユーザは友人知人関係以外からフォローされている割合が多いことが特徴として挙げられる。そのため、有用な情報をツイートするユーザを発見するためには、友人・知人関係以外からフォローされている量に注目することが必要である。提案手法では、友人知人関係をフォロー関係から取り除く。そして、友人以外からフォローされている関係の構造に HITS アルゴリズムを適用し、ユーザごとの有用度を求め、それをユーザの分類に利用する。

次に、もう1つの提案手法であるお気に入りを用いたものについて説明する。twitter のツイートには、個人の現状や状態を発信しているものと周りに情報を発信しているものに大きく分けることができる。お気に入り登録されているツイートは他のユーザが興味を持っているツイートであり、情報発信意識の高いツイートであることが多い。また、他のユーザに受け入れられているという点から、それを見るのが他のユーザにとって有用である可能性が高いといえる。このお気に入り登録の性質を用いて、他のユーザへの情報発信要求が高く、有用な情報をツイートするユーザとそうでないユーザを分類する。提案手法では、フォロワーからそのユーザのツイートがお気に入り登録されている割合を使って分類を行う。また、フォロワーごとのお気に入り登録の価値についても考慮する。情報収集意識が高いユーザからお気に入り登録されているツイートは価値が高いとして、分類を行う。

ユーザ分類後のツイートの分類は、Web ページ情報を含むツイートに着目することやツイートのリアルタイム性に着目することやトピック・テーマに応じた分類を行うことを考えているが、本論文ではツイート分類の前段階であるユーザ分類に関して主に述べる。

図1は提案する手法全体の流れである。まず、twitter 上からユーザのフォロー関係を抽出する。その後、2種類の提案手法でユーザの有用度を推定する。有用度が高いユーザのツイートから、さらに有用であるものとそうでないものにツイート分類を行う。そして、最終的に有用なユーザの有用なツイートが出力されるという流れである。分類手法と評価実験については、4章、5章で詳しく述べる。

2. 関連研究

twitter 上で影響力のある発言を行うユーザを、PageRank を用いてランキングを行う研究がある[1]。この研究では twitter 上での発言の広がりや全ユーザの中で他のユーザに影響のある

(注3): クエリを用いて検索することで、今その単語に関してどのようなツイートがされているかを表示してくれる機能である。

(注4): twitter はユーザの検索機能がユーザ ID 検索のみである。

発言をしたユーザを発見することに注目している．ここでの，他に影響のある発言とは，ユーザを伝うことで広い範囲に伝搬する発言のことを指す．本研究は，影響のある発言を行うユーザを発見するのではなく，有用なツイートをするユーザを発見するという点を重視しているため，この研究とは目的が異なる．また，現在注目を集めているキーワードや事象をツイートから抽出する手法がある [2]．本研究では，情報としてユーザが興味がありそうなものを，時間に関係なく，twitter 上からマイニングするという点でこの研究とは異なる．また，本研究ではリンク構造から友人関係を抽出してネットワーク解析を行うが，リンク構造から友人関係の抽出を行うことに類似のもので，blog コミュニティを発見する研究がある．これは，blog のリンク構造に PageRank を適用し，リンクの弱い部分を切り離すことでコミュニティにクラスタリングするという手法である [3]．

3. ユーザとツイートの特徴とその分類

この章では，twitter のツイートやユーザには，どのような種類があり，どのように分類すれば有益であるかについて説明する．twitter のツイートは情報の質に関して，大きく分けて以下の 4 つの属性をもつ．

- 情報発信性
- リアルタイム性
- 社会性
- 有用性

1 つ目の情報発信性は，そのツイートが他のユーザに対して発信している意識が高いかどうかである．twitter 上でやり取りされるツイートは，「眠い」や「遅刻した」など自分の状況をただ発信するだけの自己完結型のものと，「他のツイートへの返信」や「飲み会へのお誘い」など他のユーザが見ることを意識した発信に分けることができる．また，ツイートの情報発信性はこのようなコミュニケーションのツイートのみに限らず，もっと広い範囲に向けて，情報やニュースを発信するようなツイートもある．web ページの URL を内容に含むツイートは，内容が情報やニュースである場合が多いので情報発信性が高いといえる．また，URL を含まない場合でも，情報やニュースは情報発信性の高いツイートである．その場合は他のツイートと比べてリツイート^(注5)が多くなされるといった傾向がある．情報発信性の高いツイートは他のユーザが見ることが有用である場合が多いので，以上の傾向を用いて情報発信性をもとにツイートを分類することは有益である．

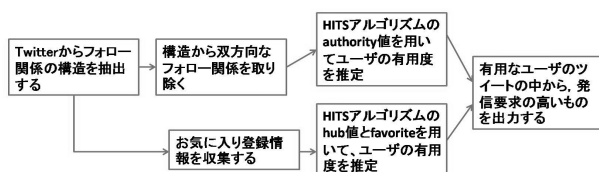


図 1 提案手法の流れ

(注5)：他のユーザーのツイートを再投稿、または引用して投稿することである．

2 つ目のリアルタイム性は，そのツイートをリアルタイムに見ることが有益であるかを表すものである．ツイートにはツイートされた時点で見ると，価値が大きくなるものと，価値が時間に左右されないものがある．例えば，地震情報などのツイートであれば，その瞬間に見ることに意味がある．また，過去の歴史に関する知識のツイートなどであれば，その瞬間に見ることに価値があるわけではない．リアルタイム性にもとづくツイートの分類の応用としては，リアルタイム検索の結果にリアルタイム性の高いツイートを提示するなどが考えられる．

3 つ目の社会性とは，一般に広く有用であるかどうかである．ツイートの内容にはニュース記事のような一般的な話題もあれば，専門家にしか理解できないような話題もある．誰が見ても理解できるようなツイートは社会性が高いといえる．逆に，その分野に精通した一部の専門家にしか理解できないようなツイートは社会性が低いといえる．社会性に関しての分類は，ツイートごとのトピック分類や一般性を分類することが考えられる．

4 つ目の有用性とは，ツイートが他のユーザにとって，有用であるか否かの指標である．有用なツイートとそうではないツイートを分類することは有益である．また，ツイートの有用性には，「情報」としての有用性と「ネタ」としての有用性の 2 種類ある．情報として有用性のあるツイートは，内容が客観的な事実や事象に対して述べてあるものであるといえる．例えば，内容に興味のある Web ページの URL が仕込まれているものや，マスメディアが行う報道のようなものや，まだ知られていない事実や事象を投稿しているものは情報としてのツイートである．blog やニュースサイトで発信されているものも情報である．また，ネタとして有用なツイートとは，主観的で内容に笑いの要素や機知な要素が仕込まれたもののことをいう．2ch の掲示板でやりとりされるような内容はネタ性の高い投稿である場合が多い．例えば，「布団のポテンシャルが高すぎてトンネル効果を使わないと出られない」というツイートはネタであるといえる．このツイートでは布団から出られない感覚をトンネル効果を用いて比喻している．ユーザが布団の心地よさを主観的に感じ，それを内容として機知に盛り込んでいる．このように，ある事実や考えを自分の捉え方で内容に盛り込んだものはネタであるといえる．

有用なツイートがネタであるのか，情報であるのかを分類することは意味があることである．ネタ性の高いツイートはユーザごとの嗜好によって，有用であるか否かが大きく変わる．ユーザの嗜好によっては，全く受け入れられない場合があるからである．情報性の高いツイートは，トピックごとに分類を行い情報検索の結果としてツイートを提示することができれば有益であるといえる．

また，ユーザもそれぞれ以下のような属性を持つ．

- 利用目的
- 有名度
- 嗜好
- 情報収集意識
- 情報発信意識

利用目的とは、そのユーザが何を目的として twitter を利用しているかを表すものである。twitter 上には、コミュニケーションを目的として利用している人もいれば、情報発信を目的として利用している人もいれば、情報収集を目的として利用しているユーザもいる。情報発信に関していえば、マーケティングなどの広告宣伝を目的であるユーザや blog などの個人の情報発信を行っているユーザや、bot^(注6)のように、あるテーマに関して情報を発信することを目的とするユーザなど様々である。情報収集に関していえば、ツイートから何か情報を得るための利用は情報収集を目的とした利用であるといえる。また、コミュニケーションを目的に利用している人は内輪向けのコミュニケーションのためであるか、もっと広い範囲でコミュニケーションを行うためであるかの2種類に分けられる。広い範囲でコミュニケーションを行うユーザはたくさんのユーザをフォローしている傾向がある。こういったユーザの間ではフォロー返しという見知らぬユーザにフォローされてもフォローし返すということが行われる。そのため、こういったコミュニケーションを行うユーザはフォロワー数もフォロー数も多くなる。このようにユーザの利用目的は様々である。利用目的が情報発信であり、その中でも広告や宣伝を発信することを目的としないものは有用である可能性が高い。そのため、この広告宣伝以外の情報発信を行うユーザを抽出することは有用である。

有名度とは、現実世界や web 上でのそのユーザの人気のことである。ユーザは有名度に基づきフォロワー数が異なるという特徴がある。例えば、芸能人、政治家などは有名なユーザであり、人気のあるユーザといえる。その他にも、会社経営を行っていたり、有名ブロガーであるといった現実世界ではあまり人気はないが、web 上では人気が高いユーザもいる。このような有名度の高いユーザは、そのユーザが発信する情報の質や情報発信性に関係なくフォロワー数が多くなる。そして、人気が高ければ高いほど、そのユーザが発信するツイートは興味があるユーザにとって有用である。

嗜好はそれぞれのユーザごとにあり、そのツイートやユーザを有用と感じるか感じないかを表すものである。ツイートに関して言えば、そのツイートが自分の興味のあるものであれば有用と感じる。ユーザに関して言えば、実世界の関係であったり、その人の有名度によって有用か否かが決まる。ユーザが持つ嗜好は情報を発信する傾向にも影響がある。例えば、車に興味を持っているユーザは、車に関するツイートをするユーザをフォローするだけでなく、自らも車に関して、ツイートしている場合が多い。このように、ユーザの嗜好によって、発信するツイートの傾向やフォローする傾向に違いがある。そのユーザが持つ嗜好に応じてユーザ分類を行うことで、そのユーザの発信・受信するトピックの傾向や属する集団を把握することができる。ユーザの嗜好するトピックを知ることによって、ツイートの検索やユーザ検索を効果的に行うことができるようになる。

情報収集意識とは、twitter 上でそのユーザが情報を収集する意識を表している。情報収集意識が高いユーザは自分の興味のあるユーザをフォローしている傾向がある。情報発信意識とは、そのユーザがフォロワーに対して情報を発信する意識の高さである。情報発信意識が高いユーザであるほど、他のユーザに見られる意識を持ってツイートを発信する。そのため、情報発信意識の高いユーザであるほど、ツイートが有用であったり、面白かったりする。情報発信意識の高いユーザとそうでないユーザを分類して、有用なユーザのツイートを抽出することは意味のあることである。

今回の実験では、有用なツイートを抽出する前段階として、ユーザの分類を行う手法を提案する。情報発信意識とユーザの有用性の間に相関があると考えられることから、情報発信意識をもとにユーザを分類する。以下の章で、提案手法について説明していく。

4. フォロー関係からの有用なユーザの推定

この章では、友人関係を取り除いたリンク構造から、有用なユーザを推定する手法について説明する。ここでいう、友人関係の定義は、お互いにフォローしあっている場合である。また、ユーザとあるユーザの間には、フォローされる側とフォロワーという関係がある。今回の実験では、フォロワーからあるユーザにされるフォローを web でいうところのリンクに見立て解析を行う。まず、twitter 上のフォロー関係から友人関係を取り除いた構造の抽出を行う。そして、この双方向なフォロー関係を取り除いた構造に、HITS アルゴリズムを適用する。HITS アルゴリズムは「authority」と「hub」という、Web ページの有用性を表す二つの尺度を用いていることが特徴的である。ページ p からページ q へリンクされていることを、 $q \rightarrow p$ と表すとすると、ページ p の「authority」と「hub」は以下のように定義される [4]。

$$auth(p) = \sum_{\{q|q \rightarrow p\}} hub(q)$$

$$hub(p) = \sum_{\{q|p \rightarrow q\}} auth(q)$$

twitter のユーザは情報発信意識と情報収集意識を持っており、それらの度合いを HITS の「authority 度」と「hub 度」で表すことにする。情報発信意識が高く、情報収集意識の高いユーザから受け入れられているユーザを authority ユーザとし、また、よりよい情報を得ようと質の高い情報を収集する意識が高いユーザを hub ユーザということにする。ユーザの authority 度は情報収集意識の高い hub ユーザから、多く参照されていると高くなり、hub 値は情報発信意識の高い authority ユーザを多く参照しているとき高くなる。また、友人関係を取り除いたフォロー関係上での authority 度の高いユーザは、友人同士のコミュニティ関係以外への情報発信意識が高いユーザであるといえる。このことから、このアルゴリズムでは情報発信意識の高い有用なユーザである authority ユーザを最終的に抽出する。

(注6): 自動で twitter に投稿するプログラム、RSS など配信されているデータをそのまま twitter へ転載するものや一定期間に管理者がランダムにメッセージを送るものなどがある。

友人関係の抽出と HITS アルゴリズムの適用を以下のような手順で実行する．

手順 1: 全体のフォロー関係を抽出する

手順 2: 友人関係を取り除いたフォロー関係を求める．

手順 3: 手順 2 の結果をもとに解析を行い, authority 値と hub 値を求める．

手順 4: authority ユーザを抽出する．

さらに, アルゴリズムによって抽出されたユーザが本当に有用であるかの評価を, そのユーザのツイートが有用であるか否かを評価することで行う．また, このアルゴリズムが有用であることを評価するために, 全体のフォロー関係にそのまま HITS アルゴリズムを適用した場合と比較する．評価に関しては 6 章で詳しく述べる．

5. お気に入りを用いた有用なユーザの推定

この章では, それぞれのツイートにつけられたお気に入りをもとに, 有用なユーザを推定する手法について説明する．お気に入りを用いた分類では, フォロワーからツイートにお気に入り付けられている量とそのユーザの総ツイート数との割合を有用度として用いる．また, フォロワーがつけるユーザへのお気に入りも一様な値で評価せず, よい hub ユーザから付けられたお気に入りは有用であるとする．より具体的には, あるユーザ x の有用度 $A(x)$ を以下のように定義する．

$$A(x) = \frac{1}{n} \sum_{i=1}^n \sum_{p \in P} W_p^i \cdot \text{hub}(p)$$
$$\begin{cases} W_p^i = 1 & \text{if } p \text{ が } i \text{ をお気に入り登録している} \\ = 0 & \text{otherwise} \end{cases}$$

あるユーザ x の有用度 $A(x)$ を求めることを考える．ここで, P は twitter 上から収集した全てのユーザの集合である．フォロワー p がツイート i をお気に入りに登録している場合は, $W_p^i = 1$, そうでない場合は, $W_p^i = 0$ とする． n はそのユーザ x がツイートした総ツイート数である．ツイートの価値がお気に入りを付けたユーザのハブ値の総和であるとする, あるユーザの有用度 $A(x)$ はユーザのすべてのツイートの価値の総和を総ツイート数で割ったものになる．全体のユーザ有用度 $A(x)$ を求める手順を以下に示す．

手順 1: 3 章と同じ手順で全てのユーザの hub 値を得る．

手順 2: それぞれのユーザのお気に入り登録情報を収集する．

手順 3: ユーザの hub 値とお気に入り登録情報をもとにそれぞれのユーザの有用度 $A(x)$ を求める．

3 章の実験と同様に, そのユーザのツイートが有用であるか否かでユーザの有用度評価を行う．また, hub 値を考慮することの効果測定のために, hub 値をそのツイートの価値として用いる手法と用いない手法で比較評価を行う．次章で評価について詳しく述べる．

6. 実験評価

この章では実験とその評価について述べる．あるユーザを

twitter 上からランダムに選び, そのユーザを起点にお気に入り情報とユーザのフォロー関係をクロールした．起点のユーザからフォロー関係をたどり, 到達した 6234 人のユーザについて, そのユーザからの全てのフォロー関係を収集した．また, お気に入り情報のクロールは 28000 人のユーザのお気に入り情報を, 先ほどと同様にフォロー関係からたどって収集した．その中から, フォロー関係でクロールしたユーザ 6234 人に対して付けられているお気に入り情報のみを考慮した．28000 人のユーザが対象の 6234 人にお気に入りを付けている人数は 2774 人で, 対象となるユーザに付けられているお気に入りの総件数は 8745 件であった．

6.1 フォロー関係からの有用なユーザの推定

友人構造を取り除かない場合と取り除いた場合で比較評価を行った．hits アルゴリズムで有効であると考えられる上位 100 件のユーザを対象に評価を行う．その際に, 両者の手法の違いの部分だけの評価するために, 両方の手法で 100 位以内に表れている 27 件のユーザは取り除き, 残りの重ならない 73 件のユーザの中から 30 件を選んだ．その選んだ 30 件のユーザが有用であるか否かを評価した．表 1 は実験の結果である．友人関係を取り除く前のネットワークに hits を適用した場合は, 20%の確率で有用なユーザが抽出できているのに対して, 友人関係を取り除いた後の結果では 40%の確率で有用なユーザを抽出できている．実験結果から, 友人関係を取り除いた場合のほうが有用なユーザを発見できていることがわかる．

また, それぞれの手法で抽出されたユーザは有用であるか否かの違いだけでなく, 上位にランキングされるユーザの傾向には利用目的に関して違いがある．友人関係を取り除かない手法の有用でないユーザの利用目的の内訳を表 2 に示す．フォロー返しをしてフォロー数とフォロワー数を増やしているユーザ, IM のようにコミュニケーションを目的に使っているユーザ, 情報収集や個人的な情報発信のユーザなどである．このような利用の仕方をしているユーザは, 他のユーザにとって有用である場合が少ない．また, このようなユーザは有用なツイートをする確率も低いのである．評価者によって有用でないと判定されたユーザは利用の仕方の違いはあるが, 一様に非有用なユーザである．

しかし, 表 3 の友人関係を取り除く手法の有用ではないユーザの内訳は, 有名人が 13 件, 特定の情報のみをツイートするユーザが 3 件いる．それ以外は, 個人的な情報発信を行うユーザとフォロー返しを行うユーザが 1 件ずつである．有名人や情報系のユーザは前述したように, 人によっては有用でないと判断されるが, あるユーザにとっては有用な場合がある．この手法で取りだされたユーザで, 完全にコミュニケーションなツイートや個人的で周りに有用でないツイートをしているユーザは 2 件しかない．つまり, 友人関係を取り除く手法では, 30 件中の 28 件は意味のある有益なユーザを抽出できたことになる．提案手法を用いることで, 上位のユーザからフォロー返しを行うユーザやコミュニケーションをメインに行うユーザなど, 非有用なユーザを下位にランキングすることができ, 上位に有用なユーザをランキングすることができた．このことから, 友人

関係を取り除いてネットワーク解析を行うことは有益であるといえる。

表 1 フォロー関係を用いた各手法の精度

	有用なユーザ数	精度
友人関係を取り除く前	6 人	20%
友人関係を取り除いた後	12 人	40%

表 2 友人関係を取り除かない手法によって抽出される非有用ユーザの利用傾向の内訳

有用でないと判定されたユーザの種類	人数
フォロー返しを行う	5 人
IM のような使い方をする	4 人
宣伝目的	5 人
情報収集のため利用している	6 人
個人的な情報をツイートする	4 人

表 3 友人関係を取り除く手法によって抽出される非有用ユーザの利用傾向の内訳

有用でないと判定されたユーザの種類	人数
特定の情報をツイートする	3 人
フォロー返しを行う	1 人
有名人	13 人
個人的なツイートをを行う	1 人

6.2 お気に入りを用いた有用なユーザの推定

「お気に入り」を用いて有用なユーザを抽出する手法について、実験評価を行った。hub 値を用いた場合と用いない場合で比較評価を行う。それぞれの手法で提示される上位 100 件のユーザの中から、ランダムで 30 件のユーザを選び、有用であるか否かを評価した。実験結果を表 4 に示す。両者の手法とも 10 人の有用なユーザを抽出することができており、精度は 33.3%である。友人関係を取り除いて hits アルゴリズムを適用する手法よりも精度は低い数値であることがいえる。

また、評価したユーザが発信する情報の傾向を表 5 と表 6 に示す。前述したように有用なツイートには、ネタ性が高く有用であるツイートと情報として有用なツイートの 2 種類がある。この手法で抽出される有用なユーザの傾向は、先ほどのフォロー関係を用いて有用なユーザを抽出する手法と違いがある。フォロー関係を用いてユーザを抽出した場合は、上位に有名人と情報をツイートをするユーザが位置する傾向があった。しかし、このお気に入りを用いた手法では、ネタや自分の考えをツイートするユーザと情報として有用なものをツイートするユーザが上位に位置する傾向がある。これはお気に入りを付けているツイートがネタとして有用であるものと情報として有用であるものの 2 種類があるためであり、それをもとに分類を行っているからであると考えられる。このように 2 種類の手法で抽出されるユーザの傾向にはそれぞれ違いがある。

また、表 7 と表 8 の非有用なユーザの内訳には、コミュニケーションを主な目的として利用しているユーザの他に、ネタ

をツイートするユーザがいる。情報をツイートしているユーザは評価者に有用と判定されているのに対して、ネタをツイートするユーザは非有用と評価者に判定されている場合が多い。このことから、内容が情報のツイートがお気に入り登録されている場合は広く有用である場合が多いが、ネタのツイートはユーザの嗜好によっては理解されない場合があり、広く有用ではないといえる。この実験ではネタ的なツイートをするユーザが有用ではないと判定された割合が多いため、精度が 33.3%と低い数値になっていることが考えられる。

ネタのツイートは人によって有用であるか否かの感じ方が違うので、お気に入り登録されているものを一様に評価してはいけない。また、ネタをツイートするユーザはいつも他のユーザに対して有用なネタを発信するわけではない。そのため、お気に入りされている割合がユーザの有用性を単純に表さないということがいえる。有用なネタをツイートするから、有用なユーザであるという因果関係は成立しないことが実験からわかる。このことから、ネタ的に有用なユーザを抽出することが難しいことから、ネタに関してはユーザの嗜好性を考慮したり、ツイートの内容に強く着目して分類することが必要であるといえる。

情報のツイートでお気に入り登録されているものに関して言えば、情報のツイートを多くお気に入りされているユーザは有用であるといえる。また、お気に入り登録されている情報のツイートは有用であるため、お気に入り登録されているツイートをネタであるか情報であるか分類することで、有用なツイートを発見することができる。さらに、「お気に入り」を用いた手法を情報のツイートのみで行うことで、より高い精度で有用なユーザを抽出することができると考えられる。また、両者の手法で実験結果に差異を見受けることができないので、hub 値を用いる手法と用いない手法とに違いはなく、お気に入りされている割合が重要であるといえる。

表 4 各手法における精度

	有用なユーザ数	精度
hub 値を用いた手法	10 人	33.3%
hub 値を用いない手法	10 人	33.3%

表 5 hub 値を用いるお気に入りによる有用なユーザの抽出

抽出されたユーザの内訳	人数
有用な情報を発信	5 人
有用なネタを発信	5 人
非有用な発信	20 人

表 6 hub 値を用いないお気に入りによる有用なユーザの抽出

抽出されたユーザの内訳	人数
有用な情報を発信	6 人
有用なネタを発信	4 人
非有用な発信	20 人

表 7 hub 値を用いるお気に入り手法によって抽出される非有用ユーザの利用傾向の内訳

有用でないと判定されたユーザの種類	人数
ネタをツイートする	5 人
コミュニケーションを行う	8 人
個人的なツイートを行う	7 人

表 8 hub 値を用いないお気に入り手法によって抽出される非有用ユーザの利用傾向の内訳

有用でないと判定されたユーザの種類	人数
ネタをツイートする	2 人
コミュニケーションを行う	3 人
個人的なツイートを行う	16 人

7. ま と め

本研究では、twitter のユーザやツイートに関して様々な分類があることを示し、それにもとづいて分類を行うことが有益であることを述べた。さらに、今回は 2 種類のユーザ分類の手法について実験考察を行った。双方向なフォロー関係を取り除きネットワーク解析をする手法と、そのユーザのツイートが「お気に入り」に登録される頻度を用いる手法である。双方向なフォロー関係を取り除いてネットワーク解析を行うことで、有用なユーザを抽出し、分類できることを示した。また、「お気に入り」に登録される頻度を用いる手法においては、ネタとして有用なユーザと情報として有用なユーザを抽出することができることを示した。また、2 種類の手法で抽出される有用なユーザの傾向の違いについて確認した。

今後は 2 つの手法を組み合わせ、情報として有用なツイートをするユーザを発見する手法を評価していく。また、2 章で述べたように、twitter のユーザとツイートには様々な分類が考えられる。ツイートの検索においては、そのツイートのリアルタイム性や社会性にもとづいて分類を行うことも重要であると考えている。本論文で述べた様々な分類手法を組み合わせることで、twitter のツイート検索システムやユーザ検索システムを実装していくことが課題である。

文 献

[1] J. J. Jianshu Weng, Ee-peng Lim and Q. He: “Twitter-rank: Finding topic-sensitive influential twitterers”, Web Information and Data Management (2010).

[2] A. P. S. A. R. R. M. Meenakshi Nagarajan, Karthik Gomadam and A. Jadhav: “Spatio-temporal-thematic analysis of citizen sensor data: Challenges and experiences”, Web Information Systems Engineering (2009).

[3] 谷口, 松尾, 石塚: “Blog コミュニティの抽出と分析”, 第 6 回セマンティックウェブとオントロジー研究会 (2004).

[4] 小山 聡 田中 克己: “Web 検索とクエリログを併用した同位語発見手法”, 日本データベース学会 Letters, 5, 4, pp. 37–40 (2007).

[5] D. G. J. K. R. K. P. R. S. R. Soumen Chakrabarti, Byron E. Dom and A. S. Tomkins: “Mining the link structure of the world wide web”, IEEE Computer (1999).

[6] 横山昌平 石川博: “マイクロブログを対象とした制約付きクラスタリングシステムの試作”, 楽天研究開発シンポジウム (2009).

[7] Y.-R. Lin, H. Sundaram, Y. Chi, J. Tatemura, B. L. Tseng : “Blog community discovery and evolution based on mutual

awareness expansion”, Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence, pp. 48–56 (2007).