O-034

# Twitterのフォロー関係による分類とその集団の可視化

Classification and Visualization of Member Network in Twitter

後迫 康宏 †　大久保 諒 ‡　森井 昌克 †
Yasuhiro Ushirozako　Ryo Okubo　Masakatu Morii

## 1 はじめに

近年，Twitter[1] のユーザ数が増加している．ユーザは趣味に関する情報や専門的な技術に関する情報など多種多様な情報を Twitter に投稿する．現在はその膨大な情報から有用な情報を抽出することを目的としたシステムが盛んに研究開発されている．Twitter から情報を抽出するシステムとして 1) 話題になっている言葉を抽出する buzztter[2] や 2) 地震に関する発言を抽出するつくるウェブ EarthquakeNotifier[3]，3) 発言内容からユーザを結びつけることばで結ぶ君 [4] などがある．1) のシステムは現在のトレンドを抽出する目的で利用される．2) のシステムは最新情報をリアルタイムに得る目的で利用される．3) のシステムはユーザを分類する目的で利用される．本研究では特に対象のユーザ（以下，起点ユーザと称する）に対して興味をもっているユーザがどのような属性を持つかを調べることを目的として 3) の方法を提案する．ここで属性とはユーザの職業や趣味に関して興味を持っている分野を指す．

Twitter ではユーザ毎に任意の内容でプロフィールを作成することができ，興味のあるユーザをフォローすることでつながりをもつことができる．発言内容からユーザを分類する方法では同一の単語を発言しているユーザには関連性があるとしてユーザを分類する．しかしユーザの投稿した情報やプロフィールは必ずしもユーザの属性を反映していない．投稿した情報やプロフィール以外の情報からユーザの属性を抽出する方法が求められている．

本研究では Twitter に対しグラフ理論を適用することでフォロー関係の情報からユーザを集団に分類する方式を提案する．具体的にはまず起点ユーザと各ユーザの関連性の強さを定義する．あるユーザのフォロワーのうち，起点ユーザと共通のフォロワーが多く含まれるほど関連性が強くなる．関係性の強さがある閾値以上のユーザを対象として Kamada-Kawai 法 [5] を用いて各ユーザに座標を与えることでクラスタに分類する．各クラスタに属するユーザのプロフィール情報から共通の属性を求め，そのクラスタがどういったユーザによるクラスタであるかを求める．一般的な方法と異なり，提案方式では数人のユーザのみ正確なプロフィール情報を作成していればどのようなユーザによるクラスタかを求めることができる．

## 2 Twitter とそのネットワーク

### 2.1 Twitter

Twitter とは発信したいことを 140 文字以内の文字列にして投稿するブログのようなものである．ユーザ同士はフォローという関係でつながっており，あるユーザをフォローすればそのユーザの発言がホームページに表示されるようになる．Twitter 以外のソーシャルネットワーキングサービスでは他のユーザとつながりを持つためには相手の了承も必要であるが Twitter ではその必要がない．つまり好きな有名人や興味のある分野に精通した人などもフォローすることができる．ここであるユーザからみてフォローを受けているユーザをフォロワー，フォローしているユーザをフレンドという．フォローの特性上フレンドに当たるユーザ同士がどのような関係を持つかは把握できるが，フォロワーに関してはそれが困難である．

### 2.2 ネットワークのモデル化

本研究ではあるユーザのフォロワー関係ネットワークを見ることによってフォロワー同士がどのような関係性を持つかを求める．フォロワー関係のネットワークを根付き木としてモデル化する．基本的な考え方は以下の通りである．図 1 において起点ユーザを $n_{0,0}$ とし，そのフォロワーを深さ 1 のユーザ $n_{1,i}$,$i$ は任意の 0 を含む正の整数とする．深さ 1 のユーザがすべてそろった時点でそれらのユーザのフォロワーを深さ 2 のユーザ $n_{2,i}$ とする．以降同様にしてユーザを取得しその深さを決定する．最終的にある深さを設定しそれ以降のユーザを取得しないことによって木が完成する．完成した木はその経路によって集団に分類できると考えられる．つまり初めに設定したユーザのフォロワー同士の関係性が導き出せると推測できる．

## 3 フォロワー関係に基づく分類

### 3.1 提案方式の概要

起点ユーザの周辺に存在するユーザを集団に分類する手法を提案する．具体的にはまず解析対象として選択したユーザとそのフォロワーとの関係の強さを表す指標を導入する．次に導入した指標を用いて分類対象の絞り込

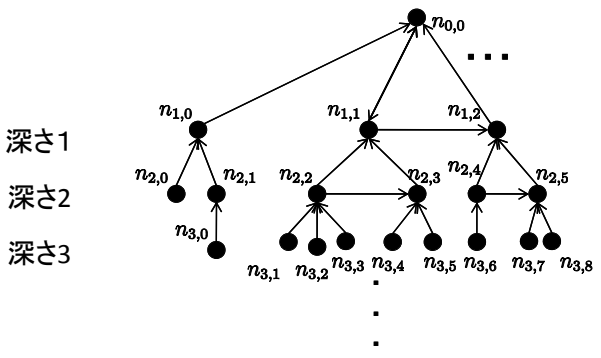


図 1　Twitter におけるネットワーク

† 神戸大学大学院工学研究科，Graduate School of Engineering, Kobe University
‡ 神戸大学工学部，Faculty of Engineering, Kobe University

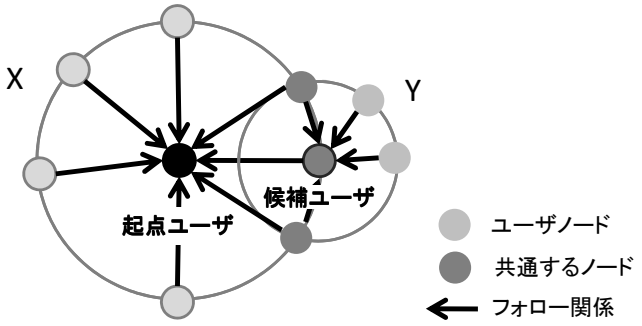


図 2　関係性の概念図

む．Twitter では興味があるユーザをフォローすることからユーザ間の距離が 1 のとき，同じ属性を持つ確率が高く，距離が離れるほど同じ属性を持つ確率が低くなると考えることができる．そこで起点ユーザと各ユーザの距離が近いほど近い座標 $(x, y, z)$ を与え，遠いほど遠くの座標を与えることができる KK 法を用いる．関連性が強いユーザほど近くの座標に配置されることになり分類が可能となる．

### 3.2　分類対象の絞り込み

条件付けにより分類対象を絞り込み，起点ユーザとの関係が強いユーザを抽出する．ここで起点ユーザとは，解析対象として最初に選択するユーザのことである．

起点ユーザとの関係の強さを「関係性」と呼ぶ．関係性は閾値付き Jaccard 係数を用いて決定する．閾値付き Jaccard 係数とはある二つの集合 $X, Y$ の類似度を求める方法で以下の式で表される [6].

$$J_{th}(X,Y) = \begin{cases} \dfrac{|X \cap Y|}{|X \cup Y|} & (|X| > k, |Y| > k) \\ 0 & (otherwise) \end{cases} \tag{1}$$

$X, Y$ の要素数 $|X \cup Y|$ が一定の閾値 $k$ に達しない場合には類似度 $J_{th}(X, Y)$ を 0 とする．提案方式では $X$ を起点ユーザと関係性が強いユーザの集合，$Y$ を候補ユーザのフォロワーの集合とする．式 (1) は候補ユーザを変更するたびに $X$ の値が大きくなってしまう．この影響を小さくするため提案方式では類似度を関係性 $R(X, Y)$ として扱う．ここで候補ユーザとは起点ユーザとの関係が強いかどうかの判定を行いたいユーザのことである．関係性 $R(X, Y)$ は次式で求められる．

$$R(X,Y) = \begin{cases} \dfrac{|X \cap Y|}{|Y|} & (|Y| > k) \\ 0 & (otherwise) \end{cases} \tag{2}$$

例として図 2 の場合について候補ユーザの関係性 $R(X, Y)$ を求める．候補ユーザのフォロワーは右側の円上にある四人である．そのうち，左側の円であらわされる集合にも属するフォロワーは濃い灰色で示した二人である．よって式 (2) よりこの候補ユーザの関係性は $2/4 = 0.5$ と求められる．

各ユーザに対して関係性を計算し，あらかじめ設定した閾値と比較することで関係の強弱を判定する．関係性が閾値を下回った場合は，起点ユーザとの関係が弱いと判断して起点ユーザが属する集合 $X$ から除外する．一方関係性が閾値を上回った場合は，起点ユーザとの関係が強いと判断し除外しない．ひととおり全てのフォロワーに対して関係性による判定と除外を行った後，関係性の強いユーザのフォロワーを辿りさらに関係性による判定を行う．以上の手順を規定数回繰り返すことでネットワークを成長させていき，起点ユーザとの関係が強いユーザにより構成されるネットワークを抽出する．ここで，$n$ 回目の繰り返しで関係性が強いと判断したユーザを第 $n$ 世代に属するユーザ，最大の繰り返し回数を最大世代数と称する．

### 3.3　Kamada-Kawai 法による分類

KK 法の基本的な考え方は，任意の 2 頂点についてグラフ理論における最短パス長が短い頂点同士は近くに配置し，最短パス長が長い頂点同士は遠くに配置するというものである．具体的にはグラフ全体に対してエネルギーを定義し，そのエネルギーが最小となるような頂点の配置を行う．グラフ $G(= \{n_{0,0}, n_{0,1}, \cdots\})$ 全体のエネルギー $E$ は次式で定義する．

$$E = \sum_{i<j} \frac{1}{2} k_{i,j} (r_{i,j} - l_{i,j})^2 \tag{3}$$

ここで，$r_{i,j}$ は頂点 $i, j (\in G)$ 間の描画上の距離，$l_{i,j}$ は頂点 $i, j$ 間の理想距離である．理想距離 $l_{i,j}$ とは頂点 $i, j$ 間の距離の目標値であり，2 頂点間の $G$ 上における最短パス長から求める．$k_{i,j}$ は $l_{i,j}^{\alpha}$ に反比例した値を持つ定数であり，次式のとおりである．

$$k_{i,j} = l_{i,j}^{-\alpha} \tag{4}$$

$\alpha$ は任意の値を与える．

KK 法は無向グラフのレイアウトアルゴリズムであるため，Twitter のフォロー関係をはじめとした有向グラフにそのまま適用することはできない．そこで便宜的にフォロー関係を無向グラフとみなすことで KK 法の適用が可能となる．しかし相互の関係は一方の関係よりも強く，それらを同列に考えることは好ましくない．そこで KK 法に相互フォローを考慮した改良を加えた手法を用いる．具体的には，ユーザ同士が相互フォローでつながっている場合に理想距離を定数で割ることで，一方通行のフォローでつながっているユーザ同士よりも近くに配置する．

### 3.4　提案方式のアルゴリズム

分類対象の絞り込みアルゴリズムを Algorithm1 に示す．入力として Twitter のフォロー関係ネットワークおよび起点ユーザ，関係性の閾値，フォロワー数の最低閾値，最大世代数を与えると，起点ユーザとの関係が強いユーザの集合を出力する．

## 4　可視化とその評価

本章ではあるユーザ A 氏および B 氏の Twitter アカウントに対して提案手法による解析を行い，その結果について評価を行う．

**Algorithm 1** 分類対象の絞り込みアルゴリズム

**Require:** Twitter のネットワーク $N = \{n_{i,j}\}$ および起点ユーザ $n_{0,0}$, 関係性の閾値 $r_{th}$, フォロワー数の最低閾値 $k_{min}$, 最大世代数 $i_{max}$

**Ensure:** 起点ユーザとの関係が強いユーザの集合 $M$

x, y := 任意の非負整数
List := 配列
$F_{n_{l,m}} := \{n_{x,y} \mid n_{x,y} \in N, n_{x,y}$は $n_{l,m}$のフォロワー $\}$
$M_0 := \{n_{0,0}\}$
$M_1 := F_{n_{0,0}}$
$M := M_0 \cup M_1$
**for** $i$ from 1 to $i_{max}$ **do**
  **for all** $n_{i,j} \in M_i$ **do**
    **if** $|M| > k_{min}$ and $|F_{n_{i,j}}| > k_{min}$ **then**
      $r_{n_{i,j}} = \dfrac{|M \cap F_{n_{i,j}}|}{|F_{n_{i,j}}|}$
    **else**
      $r_{n_{i,j}} = 0$
    **end if**
    **if** $r_{n_{i,j}} < r_{th}$ **then**
      add $n_{i,j}$ to List
    **else**
      add $n_{x,y} \in F_{n_{i,j}}$ to $M_{i+1}$
    **end if**
  **end for**
  delete $n_{x,y} \in$ List from $M$
  add $n_{x,y} \in M_{i+1}$ to $M$
**end for**

### 4.1 実験に与えるパラメータ

**関係性の閾値 $r_{th}$**
全体で 100〜200 人程度のユーザが抽出される値に設定する．適した値はユーザごとに変化するので，数度試行を行ったうえで適した値を選ぶ．

**最大世代数 $i_{max}$**
全評価実験で 5 に固定．

**フォロワー数による制限 $k_{min}$**
3 人以上のフォロワーを持つユーザを対象とする．また Twitter から取得できるデータ量に制限があるため，解析の対象とするユーザにフォロワー数の制限を設けることで多量のデータ取得が必要なユーザを解析対象から外す．フォロワーが平均の約 10 倍である 2000 人以上存在するユーザを除外することとした．

**用いるデータセット $N$**
2011 年 6 月 1 日以降に Twitter から取得したフォロー関係に関するデータ．

**可視化表示**
KK 法により各ユーザに与えた座標をもとに可視化をする．各ノードはユーザを，各ラインはフォロー関係を表す．可視化に際してフォロー関係の方向は考慮しない．

### 4.2 分類結果の評価

あるユーザ A 氏および B 氏のアカウントに対して提案手法による解析および可視化を行い，分類結果の評価を行う．関係性の閾値を 0.2 として A 氏のアカウントを解析した結果，121 人のユーザが得られた．また関係性の閾値を 0.3 として B 氏のアカウントを解析した結果，237 人のユーザが得られた．これらの分類を可視化したものを図 3 および図 4 に示す．可視化を行った結果から関係の強いユーザが四つのクラスタに分類出来ることが視覚的にわかる．

次に可視化上の分類には A 氏および B 氏の特徴との関連があるのかを評価する．各クラスタごとにユーザを無作為に選び，プロフィールや発言内容からどのようなユーザが分類されているのか推定を行った．結果を表 1 に示

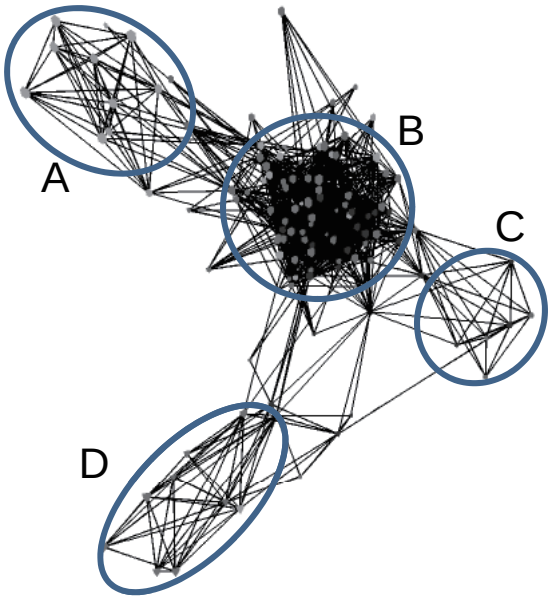


図 3　A 氏における可視化結果からの分類例

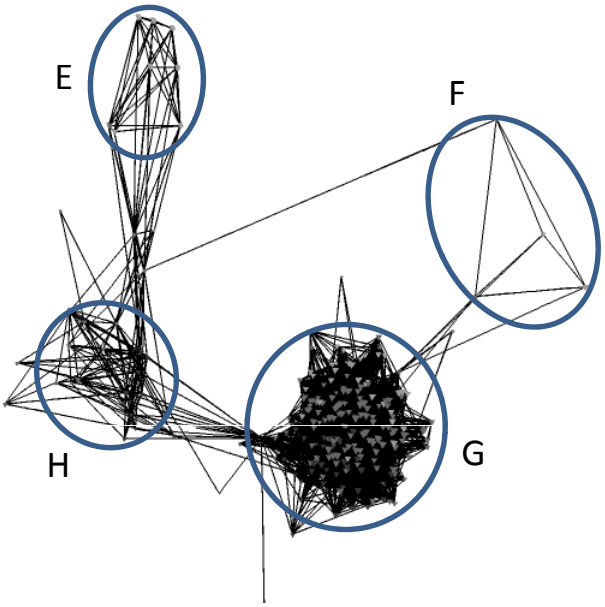


図 4　B 氏における可視化結果からの分類例

| 分類記号 | 分類の推定結果 |
|---|---|
| A | 研究室の構成員 |
| B | 特定分野の研究者 |
| C | A 氏が所属する大学の学生 |
| D | A 氏が所属する研究室の構成員 |
| E | 研究室の構成員 (分類 A と同一) |
| F | 研究室の構成員 |
| G | B 氏が所属する大学の構成員 |
| H | 特定分野の研究者 |

表 1　発言やプロフィールからの分類推定

す．それぞれのクラスタには特定の共通点を持ったユーザが多く集まっており，その共通点には A 氏および B 氏の特徴との関連がみられた．以上より，提案手法による集団の分類は解析対象として選択したユーザの特徴を反映しているといえる．

## 5　まとめ

本稿では Twitter においてつながりの情報からユーザをいくつかの集団に分類する手法の提案と評価を行った．まず分類対象を絞り込むために，起点ユーザと関連性が強いユーザを抽出した．抽出したユーザを対象に Kamada-Kawai 法を用いて座標を与えることで分類した．また数人のユーザの発言内容やプロフィールから各分類の特徴を推定すると，解析対象とした選択したユーザの特徴との関連がみられた．以上より提案方式では発言内容やプロフィールといった具体的な情報を用いることなくユーザを意味のある集団に分類することが可能だと考えられる．

## 参考文献

[1] Twitter, available at http://twitter.com/
[2] buzztter, available at http://buzztter.com/
[3] つくるウェブ EarthquakeNotifier, available at http://www.tsukuruweb.com/software/
[4] ことばで結ぶ君 α 版, available at http://twitter.taotao.ws/groups/index
[5] T. Kamada S. Kawai, “An Algorithm for Drawing General Undirected Graphs,” Information Processing Letters 31, pp.7-15, 1989.
[6] 松尾豊, 友部博教, 橋田浩一, 中島秀之, 石塚満, “Web 上の情報からの人間関係ネットワークの抽出,” 人工知能学会論文誌, 20 巻 1 号 E, pp.46-56, 2005.