Twitterにおけるユーザ同士の会話に基づいた親密度の評価と時系列的変化の可視化

小寺　暁久† 　横山　昌平†† 　山田　文康†

† 静岡大学情報学部　〒 432–8011 静岡県浜松市中区城北 3-5-1
†† 静岡大学大学院情報学研究科　〒 432–8011 静岡県浜松市中区城北 3-5-1
E-mail: †cs11032@s.inf.shizuoka.ac.jp, ††{yokoyama,fyamada}@inf.shizuoka.ac.jp

あらまし　本研究では，Twitter を対象としたユーザ同士の会話に基づく親密度を算出し，親密度の時系列的変化を可視化するシステムの構築を行う．従来のユーザがもつ影響力分析は，ユーザ間のフォロー関係のみを考慮するものであった．しかし，単にフォロー関係のみではフォローされたらフォロー仕返すユーザがいることや，フォローしているユーザごとの交友関係が密接であるか，または希薄であるかを判定することが困難であるという問題があった．そこで，本研究では，ユーザ同士の会話に着目し，ユーザ同士の親密度を算出することで，本質的な交友関係を示す事を可能にすることを目指す．提案システムでは，時間変化に伴いユーザ間の親密度が直感的に理解できるものとする．

キーワード　ソーシャルメディア，ソーシャルグラフ

1. はじめに

近年、様々なソーシャルネットワーキングサービス（SNS）が登場した．ユーザは SNS 上で情報発信や情報共有を行うことができる．ユーザが発信する情報は大量かつ膨大であり，それらを活用することはマーケティングや広告提供，企業活動などで有用とされている．

中でも代表的な SNS として Twitter(注1) がある．Twitter の特徴として，他の SNS と比べて即時性に優れている点と，一方向のみの繋がりが存在する点がある．即時性に優れている点で，ユーザが発信した情報をリアルタイムに取得することが可能である．一方向のみの繋がりが存在する点では，facebook(注2) や mixi(注3) などのようなユーザが申請を申し出て，もう一方のユーザが承認することで繋がる関係ではなく，相手の承認の可否に関わらず繋がりを持つことができる．フォロー関係を基に，ユーザが属するコミュニティの抽出や，そのユーザに関する属性情報を分析することができる．

本研究に関連する用語について説明する．本稿ではフォロー，リプライという用語を頻繁に利用する．フォローとはユーザ間の繋がりを示しており，フォローすることによりフォローしたユーザのツイートを自身のタイムラインに表示することができる．リプライとは，他のユーザによって投稿されたツイートに対して返信するツイートのことをさす．Twitter でのリプライの様子を図 1 に示す．図 1 に示すように，リプライの基となるツイートが一番上に表示され，それに対するリプライがその下に順次表示される．

現在、Twitter 上に存在する膨大なデータを活用する研究は

図 1　Twitter で行われるリプライの様子

さかんに行われている．奥村 [1] によると，Twitter を代表とするマイクロブログのデータに対して行われている研究として，Authority 分析，評判分析，実世界の動向，マイクロブログの書き手の属性推定，マイクロブログのトピック同定，トレンド分析，自動要約，情報の信頼性評価などがあると分類している．各手法においてフォロー関係によるユーザのつながりに着目したものは多く，公式 Twitter でも機能として実装されているユーザに対しておすすめのユーザを推薦する機能に関する研究 [6] では，ユーザ間のフォロー関係に着目したものが多い．また，Authority 分析とよばれるユーザがもつ影響力がどの程度あるのかを計測する分析では，ユーザのフォロワー数に着目し，ユーザがツイートによってどの程度の影響をフォロワーに対して与えるのかといった研究 [4] もある．

しかしながら，これらの研究で行われているようにユーザのフォロー関係を見てそれらに対して同じように繋がりがあると判断することについて 3 つの問題が挙げられる．1 つは Twitter ではフォローをされたらフォローを仕返すユーザもいること，2 つ目は著名人・芸能人のようにフォローネットワーク内に関係の薄いユーザが混在していること，3 つ目に一度フォローをするとそのフォローを解除するという操作があまり行われないため，疎遠になっているユーザと親密であるユーザとの区別が

(注1)：https://twitter.com/
(注2)：https://www.facebook.com/
(注3)：https://mixi.jp/

できないことがある．これらの問題から，同じフォロー関係にあるユーザでも，ユーザによって親密度の度合いに違いが存在する．

そこで，本研究ではユーザの本質的な交友関係を抽出する手法を提案し，親密度の時系列的変化を可視化することを目指す．本質的な交友関係を抽出するために，ユーザ同士の会話に基づきユーザ間の親密度を算出する．親密なユーザ同士では，短時間で多くの会話が行われていると考えられる．また，会話の開始となるリプライを送るユーザは相手と交友を深めようとしていると考えられる．そのため，提案手法では，会話数やリプライの時間間隔を用いて親密度の算出を行う．ユーザ間の親密度を算出することで，同じフォロー関係においても親密度の違いが分かる．

一般的に親密度は，時間に伴って変化すると考えられる．過去に親密であったユーザであっても，最近では，疎遠になっているユーザが存在することがその一例である．提案システムでは，特定の期間内の会話から算出された親密度を基に可視化を行う．ユーザをノードとし，フォローによるリンクをエッジとしたソーシャルグラフによって可視化することによって，直感的にどのユーザと親密であるかを理解することが可能である．

本論文では，以下の 2 章で関連研究を述べたのち，3 章で親密度の算出方法と提案システムについて述べる．4 章では親密度の有効性を実験・評価し，5 章でまとめを述べる．

2. 関 連 研 究

Twitter でのフォロー関係に基づく研究は既に行われている．Java ら [2] はユーザのフォロー関係から 3 つのグループに分類し，Kwak ら [3] はフォロー関係によって構築されるネットワークの特徴を調査した．特定のフォローネットワーク内における研究では，Cha ら [4] がユーザがもつ影響力を測定した．

ソーシャルグラフに関する研究のうち，特に本研究と強く関係しているものとして，Huberman ら [5] の研究がある．Huberman らは，Twitter におけるフォロー関係によって構築されるソーシャルグラフでは実際のユーザ同士の関係を明らかにすることはできないことを指摘している．Huberman らは，ユーザ間のリプライを用いて，フォロー関係によっては見えない潜在的なネットワークの抽出を行った．

2.1 リンク予測問題に関する研究

ユーザ推薦を考えるとき，対象ユーザのまわりで構築されるネットワークに着目した研究として Liben-Nowell ら [6] の研究がある．Liben-Nowell らは，ネットワークの構造からリンクを予測し，将来的に関係性を構築するであろうユーザの予測を行った．ネットワークの類似度が高いユーザ間には新たに関係性が構築されやすいという仮定のもとに予測している．類似度の算出にあたり以下のような指標から $Score(x, y)$ を算出した．$\Gamma\ (x)$ はノード x の隣接ノード数を示す．

- Common Neighbors

ノード x とノード y の共通隣接ノード数が多いほど，2 つのノード間にリンクが存在する可能性は高いとする指標である．

$$Score(x, y) = |\Gamma(x) \cap \Gamma(y)|$$

- Jaccard's Coefficient

ノード x とノード y の共通隣接ノード数が両者の総隣接ノードに占める割合が高いほど，2 つのノード間にリンクが存在する可能性は高いとする指標である．分子に Common Neighbors を置き，割合として算出することにより，単に共通ノード数が多いだけでは $Score(x, y)$ は高くならないことが特徴である．

$$Score(x, y) = \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x) \cup \Gamma(y)|}$$

- Adamic/Adar

ノード x とノード y の共通隣接ノードに対して重みを加え，それらの総和が多いほど，2 つのノード間にリンクが存在する可能性は高いとする指標である．CommonNeightbors と異なり，共通隣接ノードを 1 としてカウントするのではなく，それぞれの共通隣接ノードに対しても隣接ノード数を求め，それを基に算出した重みを加えている．重みはノード数が少ないほど重くなるため，友人が少ないユーザを共通隣接ノードとして持っていると $Score(x, y)$ は高くなる．

$$Score(x, y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\log |\Gamma(z)|}$$

- Preferential Attachment

ノード x とノード y の隣接ノードの積で算出されるため，隣接ノード数が多いほど $Score(x, y)$ は高くなる．多くの友人を持つ者同士は将来，関係性を構築する可能性があるという考え方である．

$$Score(x, y) = |\Gamma(x)| \cdot |\Gamma(y)|$$

2.2 対話ネットワークと投稿内容を併用した研究

岡本ら [7] は，既存研究であるグラフ構造の課題を踏まえ，ユーザ間で行われる対話によって構築されるグラフ構造を用いた Random Walk 手法と，ユーザが発言する内容の類似度を考慮した手法の 2 種類を比較した．Random Walk 手法における遷移確率は，ユーザ間のリプライ数や発言内容の類似度を考慮して重み付けを行った．ユーザ推薦における Random Walk 手法では，対象ユーザを起点として隣接ノードへの移動を繰り返すことにより，移動後に位置するノードの確率が多いノードを推薦する．岡本らはユーザ間のリプライ数は親密性を表すと考え，リプライ数が多いユーザへの遷移確率が高くなるよう計算を行った．

これらの研究では，ユーザがもつフォローネットワークをそのまま扱ったものが多い．また，数は少ないものの，ユーザ間の対話に基づいたものは単にリプライ数を扱うものであった．単にリプライ数を親密度とみなすことへの問題として，ユーザや会話によってリプライ数が多いものと少ないものが存在することが挙げられる．また，単にリプライ数だけに基づくだけでは，より精度の高いリアルな交友関係を抽出することは困難であると考えられる．そこで，本研究ではユーザ間の会話に立ち入りリプライ間隔・会話の開始ユーザ/終了ユーザに着目し親密度を算出する．

3. 提 案 手 法

本章では，本質的な交友関係を抽出した際に得られる知見を述べた後，提案手法であるユーザ同士の会話に基づく親密度の算出方法について述べる．本研究における本質的な交友関係とは，建前的な関係や一方向のみの関係を排除した仲が良いとされるユーザとの関係をさす．

3.1 本質的交友関係を抽出する意義

ユーザの推薦機能を考えたときに，対象ユーザがフォローしているユーザのうち疎遠なユーザより親密なユーザを共通してフォローしているユーザの方が，対象ユーザにとって交友関係を作りやすいと考えられる．しかし，フォロー関係のみを考慮すると，疎遠なユーザと親密なユーザを同等に扱うため，疎遠なユーザをフォローしているユーザを推薦する可能性がある．そこで，フォロー関係で親密であることや疎遠であることを考慮して親密度を算出し，本質的交友関係を抽出する．

3.2 データの取得

親密度を算出するにあたり，ユーザのツイートを網羅的に取得する必要があるため，Twitter 社が提供する Twitter API(注4) のうち REST API によってデータ取得を行う．REST API では逐次的なデータ取得ができないため，定期的に実行しデータを取得し蓄積する．API 制限により，個人で利用する Streaming API では Twitter での全データを取得することができないため，本研究では対象ユーザと相互フォロー関係にあるユーザとの親密度を算出する．Twitter 全体のデータを取得することが可能となれば，自身のアカウントを対象ユーザとするだけでなく，他人の交友関係を閲覧することも可能となる．

3.3 親密度の算出方法

本研究では，一連のリプライのやりとりを会話と定義し，リプライが 3 件以上の個人対個人の会話に対して親密度の算出を行う．親密度を算出するにあたり，ユーザ間で行われた会話からユーザからユーザへの交友姿勢を算出する．ユーザ同士の親密度の算出は以下の式によって求める．

$$F_{ij} = \sqrt{(f_{ij} \times f_{ji})}$$

ここで，F_{ij} は，ユーザ i とユーザ j 間の親密度を示しており，f_{ij} は，ユーザ i からユーザ j への交友姿勢，f_{ji} は，ユーザ j からユーザ i への交友姿勢を示している．本研究において，交友姿勢とは，交友を深めようとする姿勢であり，相手に対する好感度に類似した指標とする．相乗平均を用いることで，互いの交友姿勢が同じ値のときに最大値を得るようにし，どちらか一方の交友姿勢が低ければ低い方に値は傾く特徴を反映させている．なお，会話が一度も行われなかったユーザとの親密度は 0 とする．

ユーザ i からユーザ j への交友姿勢 f_{ij} は以下に示す式を用いて算出する．

$$f_{ij} = (1 + \log T_{ij}) \times \frac{1}{T_{ij}} \sum_{k=1}^{T_{ij}} f_{ij}^{k}$$

(注4)：https://dev.twitter.com/

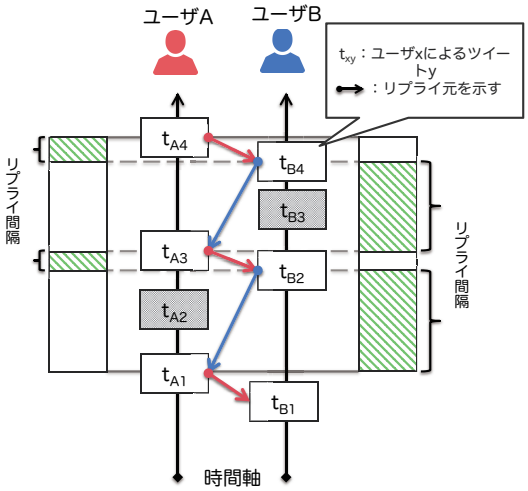

図 2　会話内交友姿勢の算出する例

ここで，T_{ij} は，ユーザ i とユーザ j 間の会話数，f_{ij}^{k} は，会話 k におけるユーザ i からユーザ j への会話内交友姿勢を表す．会話内交友姿勢は，ある会話内でのユーザ i からユーザ j への交友姿勢を表し，会話内でのリプライ間隔に基づいて算出する．

対象ユーザのリプライを起点としリプライ元を辿ることで，会話内交友姿勢を算出する．図 2 に対象ユーザ A とユーザ B との会話を時系列的に表示した例を示す．本研究において，会話内でのユーザがリプライにかかる時間をリプライ間隔とし，リプライ間隔が短いほど交友姿勢が高いという仮定のもとで行う．図 2 でのユーザの下に四角形で示されたものは 1 つのツイートであり，位置が上であるほど投稿日時が新しいものを表している．矢印が伸びるツイートはリプライであることを示す．

対象ユーザ A とユーザ B との間で行われた会話のタイムラインが図 2 のようであった場合, 対象ユーザ A の方がリプライ間隔が短いことが分かる．会話を開始したユーザは最初のリプライをしたユーザとし，ユーザ A であり，会話を終了させたユーザは最後のリプライを見送ったユーザとし，ユーザ B である．会話の起点となるツイートからの最初のリプライの時間はリプライ間隔として含まないものとする．リプライ間隔は取得したツイート情報に含まれる投稿日時から算出する．

提案手法としてリプライ間隔の他に，会話の開始ユーザ/終了ユーザに着目する．会話の開始ユーザは交友姿勢が高い，会話の終了ユーザは交友姿勢が低いと仮定し，リプライ間隔によって算出した値に開始ユーザならば自身の平均リプライ間隔を差し引き，終了ユーザならば自身の平均リプライ間隔を足すこととする．考え方として，開始ユーザによる最初のリプライは相手を待たせたという意味ではないということ，終了ユーザによる最後のリプライを見送るという行為は相手を待たせた可能性を含んでいることを意味する．会話時間に占めるリプライ間隔の割合が少ないユーザほど会話内交友姿勢が高いと評価するため，会話内交友姿勢は以下の式によって算出する．

$$f_{ij}^{k} = 1 - \frac{R\text{-}time^{k}(i)}{R\text{-}time^{k}(i) + R\text{-}time^{k}(j)}$$

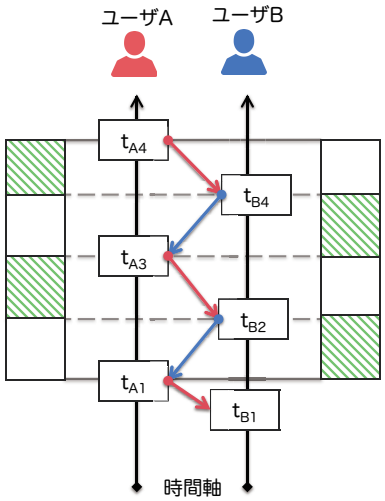

図 3　親密度が高くなるケース

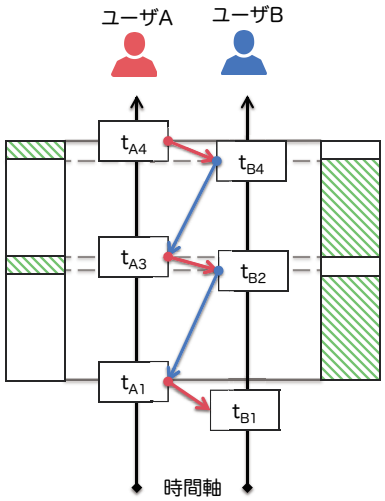

図 4　親密度が低くなるケース

$$= \frac{R\text{-}time^k(j)}{R\text{-}time^k(i) + R\text{-}time^k(j)}$$

ここで，$R\text{-}time^k(i)$ は，会話 k におけるユーザ i のリプライ間隔の総和をしめし，$R\text{-}time^k(i) + R\text{-}time^k(j)$ は会話の総時間と等しい．

3.4　親密度が高くなるケースと低くなるケース

図 3 に親密度が高くなるケースを示す．ユーザ A とユーザ B によって行われた会話が図のようなものであった場合，両者のリプライ間隔が会話全体の時間に占める割合に差がない時，両者の間の親密度は高く算出される．一方，図 4 に親密度が低くなるケースを示す．図では，ユーザ A のリプライ間隔は小さいため交友姿勢が高いのに対し，ユーザ B のリプライ間隔は大きいため交友姿勢が低いとされる．会話においてユーザのどちらか一方のみの交友姿勢が高い，もしくはどちらか一方のみの交友姿勢が低い場合に親密度は低く算出される．

3.5　提案システムの概要

提案システムでは，提案した手法によって算出された親密度を基に抽出された交友関係の可視化を行う．図 5 にシステム概要図を示す．ユーザは親密度を算出する期間をブラウザ上で設定することで，動的にデータベースから設定期間内のツイートをとりだし親密度を算出する．算出された親密度からユーザをノード，フォローリンクをエッジとしたソーシャルグラフを WebWorkers を用いて非同期処理により描写を行う．データの取り出し及び親密度の算出処理はサーバサイドで行い，グラフの描写処理はユーザサイドで行う．図 6 に提案システムによって表示される画面を示す．ユーザは特定の期間を指定することで，その期間における親密度を基に交友関係を可視化したソーシャルグラフを閲覧することができる．

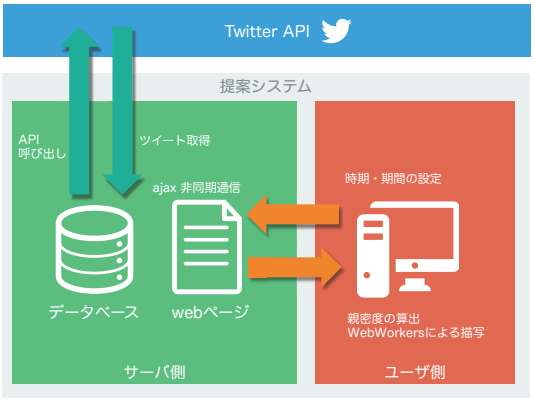

図 5　システム概要図

4.　実験・評価

本章では，3 章で述べたユーザ間の会話に基づいて算出される親密度の有効性を確認するための比較実験を行う．比較対象として，従来の研究でのリプライ数をそのまま親密度とする手法との比較を行う．

4.1　データセット

実験に用いるデータは著者アカウントから取得したデータセット 1 と，実験協力者 11 名から取得したデータセット 2 の 2 つのデータセットを準備した，それぞれのデータセットに適した実験を行った．データセット 1 は，2012 年 5 月 1 日から 2015 年 1 月 1 日までの期間で行われた著者アカウントと相互フォロー関係にあるユーザ（198 名）との間のリプライをデータセットとした．データセット 1 のうち，著者アカウントからフォローユーザへのリプライの数は 1,114 件であり，フォローユーザから著者アカウントへのリプライの数は 1,021 件であった．データセット 2 は，それぞれの実験協力者と相互フォロー関係にあるユーザから最新の 3,200 件のツイートを取得し，それらをデータセットとした．データセット 2 に含まれるユーザ数は 2,107 名であり，そのうち親密度を算出することができたユーザ数は 463 名であった．取得したツイート数は 5,182,268 件であった．

4.2　データセット 1 に対する実験

著者アカウントから取得したデータに対して，リプライ数と会話数での比較および，会話数と提案手法での比較，時系列的変化に伴う親密度の変化の 3 つの実験を行った．

4.2.1　リプライ数と会話数での比較

著者アカウントから発するリプライ数を親密度とした場合と

図 6　提案システムイメージ

表 1　リプライ数と会話数での比較

順位	リプライ数		順位	会話数	
1	A1	93	1	A1	35
2	A2	61	2	A2	23
3	A3	57	3	B2	23
4	A4	47	4	B3	22
5	A5	40	5	A5	22
6	B1	38	6	C1	22
7	B2	36	7	A4	18
8	B3	36	8	B1	17
9	B4	31	9	A3	16
10	B5	25	10	C5	12

会話数を親密度とした場合の比較を行う．表 1 はリプライ数と会話数をにユーザ別に順位化し上位 10 名を抽出したものである．本研究では 3 件以上のリプライを含む個人対個人のユーザで行われるものを会話と定義しているため，一方的なリプライのみのユーザはリプライ数での順位には表示されても会話数での順位では表示されない．データセットでのリプライを行ったユーザの数は 113 名であるのに対し，会話を行ったユーザの数は 104 名であった．表中のアルファベットと数字はユーザ名を代替する文字列であり，実際のユーザ名と対応している．ユーザ A3，ユーザ A4 はリプライ数を親密度とした場合には上位に位置しているが，会話数を親密度とした際には順位を落としていることから，1 回の会話において多くのリプライを行ったユーザと思われる．対照的に，ユーザ B2，ユーザ B3 は，会話数を親密度とした場合にリプライ数での順位より上がったことから，1 回の会話におけるリプライの回数は少ないものの，多くの会話を行ったと考えられる．このことから，会話数を親密度とすることによって，一方的なリプライを排除することができ，1 回の会話において行われるリプライ数に関わらず会話数がより多く行われたユーザが親密であるとする評価を行うことを可能とした．

4.2.2　会話数と提案手法での比較

表 2 に，3 章で述べた手法によって算出した親密度を基にユーザ別に順位化し，上位 20 名を抽出した結果を示す．カラム「ユーザへの交友姿勢」は著者アカウントからその行に対応するユーザへの交友姿勢の値を示し，カラム「ユーザからの交友姿勢」はその行に対応するユーザから著者アカウントへの交友姿勢の値を示す．相互の交友姿勢の差に 0.5 以上の差があり，交友姿勢の値が高い方に色づけを行った．表 2 において，順位と会話数を比較すると，おおよそ会話数を降順に並べ替えた時の順位と変化はないことが分かる．これは交友姿勢を算出する際に用いる会話数を対数でとった値を係数とした影響が大きいことが考えられる．しかしながら，ユーザ C1 については期間内における会話数が 22 件と比較的多いにも関わらず，22 件程の会話数を持つユーザ集団の順位と比べると順位を落としていることが分かる．著者アカウントとユーザ C1 間の交友姿勢のバランスを見ると，ユーザへの交友姿勢よりもユーザからの交友姿勢の方が値が大きく上回っていることが分かり，これはユーザ C1 から著者アカウントへの交友を深めようとする姿勢が著者アカウントからユーザ C1 への交友を深めようとする姿勢より強いことを示している．ユーザ C1 とは対照的に，ユーザ B1 については期間内での会話数が 17 件であるが，同様な会話数を持つユーザの順位と比較すると上位に位置している．著者アカウントとユーザ B1 間の交友姿勢のバランスを見ると，その差は約 0.1 と少なく交友姿勢がほぼ同等であり，その結果として親密度が高くなったことが確認できる．このことから，交友姿勢が一方的に高い場合または一方的に低い場合において，親密度は比較的低く算出されることを可能とした．

4.2.3　時系列的変化に伴う親密度の変化

親密さの程度が時間経過に伴って変化する現象を提案手法の親密度によって示すことができるか実験をおこなった．表 3 に，データ収集期間のうち 2005 年 7 月 1 日から 2015 年 1 月 1 日までの 2 年半を半年毎に分割し，それぞれの期間における親密度を算出した結果を示す．ユーザによって変動の様子は様々であるが，特に顕著なユーザとして A1，B1，C1 の 3 名があげられる．ユーザ A1 は，2013 年後期に上位 20 位より順位を落としたが，再び 2014 年前期より交友をよく行うようになったこ

表 2　提案手法による親密度の算出結果

順位	ユーザ	会話数	ユーザへの交友姿勢	ユーザからの交友姿勢	親密度
1	A1	35	1.5851	0.9590	1.2329
2	B2	23	1.0673	1.2944	1.1754
3	A5	22	0.8912	1.4512	1.1372
4	A2	23	1.5092	0.8526	1.1343
5	B3	22	1.5188	0.8236	1.1184
6	B1	17	1.0488	1.1817	1.1133
7	A4	18	1.4044	0.8509	1.0932
8	C1	22	0.7384	1.6041	1.0883
9	A3	16	0.9086	1.2955	1.0849
10	D5	12	0.9114	1.1678	1.0317
11	C2	12	1.2116	0.8676	1.0253
12	D1	11	0.9063	1.1351	1.0143
13	C5	12	1.2928	0.7864	1.0083
14	F4	10	1.2304	0.7696	0.9731
15	B4	11	1.3612	0.6802	0.9622
16	D2	8	0.9894	0.9136	0.9507
17	C4	9	1.2553	0.6990	0.9367
18	E1	8	0.7544	1.1487	0.9309
19	D4	8	1.1498	0.7532	0.9306
20	G3	8	0.7380	1.1651	0.9273

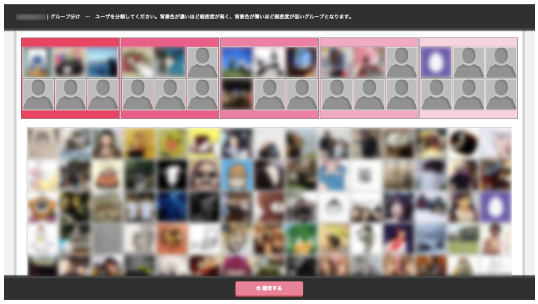

図 7　ユーザを分類する作業画面

とが分かる．ユーザ B1 は 2012 年後期と 2013 年前期で上位に位置しているが，2013 年後期以降会話が 1 回もおこなわれることがなかった．このことから，著者アカウントはユーザ B1 と以前は交友関係があったが，それ以降は疎遠になっていることが言える．またそれとは別にユーザ C1 は，各期間において上位 20 位以内に位置しているが特に高い順位の時はないことから，親密度の起伏がなく安定した交友関係を築いていることが分かる．ここに示したユーザ 3 名は実際の交友状況と親密度が深く結びついていることが確認できた．

4.3　データセット 2 に対する実験

実験協力者 11 名から取得したデータに対して，従来のリプライ数による評価と提案手法での精度の比較実験を行った．その後，得られた精度の差に対して統計解析を行った．

4.3.1　実験協力者による相互フォローユーザの分類

実験協力者には，自身と相互フォロー関係にあるユーザを仲が良いかどうかによって 5 グループに分類する作業を行ってもらった．図 7 に実際に実験協力者に行ってもらったユーザを分類する作業画面を示す．

4.3.2　親密度による適合率とリプライ数による適合率

表 4 に実験協力者が分類したユーザが属するグループを答えとして，親密度から予想されたグループとの適合率を示す．同様に表 5 には，リプライ数から予想されたグループとの適合率を示す．ここで適合率とは，実際に分類されたグループと予想されたグループが完全に一致した確率のことを言い，表では対角に示された値の総和が全体に占める比率である．表におけるグループ番号の 1 が仲が良いとされるグループであり，2 が少し仲が良いとされるグループ，3 が普通のグループ，4 がそれほど仲が良くないグループ,5 が仲が良くないグループである．それぞれの行が分類されたグループであり，列が予想されたグループであることを示す．例えば，表 4 ではグループ 1 に分類されたユーザが親密度によってグループ 3 と予想されたユーザ数が 13 人だと読み取ることができる．2 つの表から親密度による適合率は 38%で，リプライ数による適合率は 36%となり，わずかに親密度による適合率の方が高い結果が得られた．

表 4　親密度から予想されたグループとの適合率

親密度	1	2	3	4	5
1	70	44	13	5	4
2	44	63	25	13	2
3	14	27	23	19	9
4	6	10	23	17	6
5	2	3	8	8	5

表 5　リプライ数から予想されたグループとの適合率

リプライ数	1	2	3	4	5
1	89	62	25	15	4
2	68	102	50	26	10
3	27	58	46	32	14
4	8	23	43	45	20
5	3	11	13	21	19

4.3.3　相関係数の差の検定

適合率の実験では，わずかに親密度を用いた方が適合率が高い結果が得られたことから，一般的にもそのようなことが言えるのかどうかについて統計解析を用いて検証した．まず，実際に実験協力者が分類したグループ分類と親密度の相関と，グループ分類とリプライ数の相関の比較を行った．相関係数を求めた結果，グループ分類と親密度の相関係数が-0.274 で，グループ分類とリプライ数の相関係数が-0.221 であった．グループ分類は 1 が最も仲が良いとされるグループであるため，相関係数は負の値をとる．この結果から，両者とも弱い相関が見られわずかではあるが，グループ分類と親密度の方が相関が強いことが分かった．次に，この差が一般的にも言うことができるのか，2 つの相関係数の差の検定を行った．帰無仮説は，2 つの母相関係数は等しいとする．それぞれの相関係数とデータ数（463 件）から変換値を以下の式 1 によって求め，変換値より統計量を式 2 によって求めた．

$$z_i = \frac{1}{2}\ln\frac{1+r_i}{1-r_i} \tag{1}$$

$$z = \frac{z_i - z_j}{\sqrt{\frac{1}{n_i-3}+\frac{1}{n_j-3}}} \tag{2}$$

ここで，r_i は相関係数，n_i はデータ数を示す．これにより求められた統計量は-0.8599 であった．この統計量は標準正規分布に従い有意差を 5%とすると，統計量の値は帰無仮説の採択域である（-1.96 より大きく 1.96 より小さい）ため有意差は見られない．この結果から，今回実験協力者から得られたデータでは，偶然に提案手法を用いた方が従来のリプライ数を用いる手法より精度が良いとする結果が得られたが，必ずしも提案手法を用いた方が精度が良いと断言することはできない．

5.　ま　と　め

本研究では，ユーザ間の会話に基づいて親密度を算出し，本

表 3　時系列的変化に伴う親密度の変化

	2012年後期			2013年前期			2013年後期			2014年前期			2014年後期		
順位	ユーザ	会話数	親密度	ユーザ	会話数	親密度	ユーザ	会話数	親密度	ユーザ	会話数	親密度	ユーザ	会話数	親密度
1	B1	11	0.9893	B5	7	0.9088	A2	9	0.9752	A1	19	1.1071	B2	13	1.0554
2	A3	5	0.7867	A5	7	0.9075	B3	10	0.96	B2	10	0.96	A1	11	1.0046
3	D5	3	0.7384	G1	5	0.8104	A4	7	0.92	B3	9	0.9406	D4	8	0.9306
4	F4	3	0.7336	B1	4	0.784	C1	7	0.8	A2	9	0.9344	A5	7	0.9133
5	I2	3	0.7236	F3	4	0.7782	A3	5	0.8	D1	8	0.9187	C5	4	0.7955
6	C1	3	0.6	C1	6	0.7658	A5	6	0.8	C2	6	0.8751	A4	4	0.7411
7	A1	3	0.6708	C5	3	0.7335	H1	4	0.8007	D3	4	0.801	J4	3	0.7375
8	K2	2	0.6505	E1	4	0.7116	F4	4	0.7797	G2	4	0.7783	I3	3	0.7369
9	L2	2	0.6505	I5	3	0.6638	G3	4	0.7793	A3	4	0.7735	B4	3	0.7358
10	C4	2	0.642	K2	2	0.6466	D5	3	0.7369	A4	5	0.7681	N1	3	0.7348
11	H5	2	0.6229	D2	2	0.6449	D3	3	0.7328	D2	3	0.6726	A2	5	0.6521
12	C5	1	0.4899	A4	2	0.6433	J5	3	0.7236	D5	2	0.6495	H4	2	0.6505
13	U4	1	0.4899	S2	2	0.6374	G4	3	0.711	F1	2	0.641	H2	2	0.6472
14	P4	1	0.4899	E5	2	0.6374	C2	3	0.6983	C1	2	0.6374	C2	2	0.6412
15	Q1	1	0.4899	E4	2	0.6258	B4	3	0.6701	B4	4	0.6305	B3	3	0.6387
16	D1	1	0.4899	D5	2	0.6095	N3	2	0.6505	E1	2	0.6187	A3	2	0.6374
17	L5	1	0.4899	F4	2	0.962	G1	2	0.6505	H4	2	0.5858	Q5	2	0.6374
18	D2	1	0.4899	F5	2	0.554	C5	2	0.6505	G3	2	0.5858	C1	2	0.6374
19	U5	1	0.4899	H5	2	0.5709	I1	2	0.6499	A5	2	0.5378	D5	2	0.6374
20	Q2	1	0.4899	A1	2	0.5774	S4	2	0.6374	O4	2	0.5261	E4	2	0.6357

質的な交友関係を抽出し提案システムによって可視化をおこなった．従来の研究では，単にフォロー関係のみを考慮したものが多く，ユーザによって親密であることや疎遠であることなどを区別することができない問題があった．ユーザ間の会話におけるリプライ間隔・会話の開始ユーザ/終了ユーザを考慮し，ユーザからユーザへの交友姿勢を算出し交友姿勢から親密度を算出することで，ユーザによって親密であることや疎遠であることを区別することができた．

親密度の有効性について，実験協力者から取得したデータではわずかに提案手法である親密度を用いる方が，従来のリプライ数を用いる手法より精度が高い結果が得られたが，統計解析の結果より一般的に提案手法を用いる方が精度が高いと言い切ることはできなかった．

提案システムでは算出された親密度をフォローによるリンクをエッジとし重み付けを行い，ユーザをノードとするリンク構造を可視化することができた．ユーザは特定の期間を設定することで，その期間内での親密度に基づいた交友関係を閲覧することができた．また，親密度は時間変化に伴い変化することを，可視化させることで確認することができた．

今後の課題として，親密度の精度をあげることが求められる．本研究では個人対個人での会話を対象としているが，複数人による会話は考慮していないことが挙げられる．複数人による会話を考慮する場合には，リプライを 1 対 n と考え会話内交友姿勢を会話に参加するユーザで考えられる組み合わせすべてにおいて算出する．また，交友姿勢を算出するにあたり返信間隔を指標のひとつとするにあたって，講義や仕事，就寝など返信することができない時間帯を考慮する必要がある．返信することができない時間帯が存在すると，極端にそのユーザの交友姿勢が低くなるという傾向があるため，会話が行われている時間帯によって調整する処理などを加えることや，返信間隔以外の指標を取り入れるなど，より精度の高い親密度を算出することを目指す．

さいごに，提案手法の有効性を計る実験では，実験協力者が相互フォロー関係にあるユーザを分類したグループを答えとし適合率を算出したが，実験協力者自身が本質的交友関係を理解できていない可能性を含んでいる．そのため，実験協力者のみによるユーザの分類では，実験協力者はあるユーザを仲が良いグループに分類したが，そのあるユーザにも同様にグループ分類を行ってもらうと実験協力者は仲が良くないグループに分類される可能性がある．そのため，手法による精度を比較する際には，何を答えとして適合率を算出するのか慎重に行うべきだと考えられる．

文　献

[1] 奥村学，“マイクロブログマイニングの現在”，電子情報通信学会技術研究報告．NLC，言語理解とコミュニケーション，vol.111, No.427, pp.19-24, 2012.

[2] Java, A., Song, X., Finin, T., and Tseng, B., “Why We Twitter: Understanding Microblogging Usage and Communities.”, Proc. of the 9th WEBKDD and 1st SNA-KDD ‘07 workshop on Web mining and social network analysis, pp. 56-65, 2007.

[3] Kwak, H., Lee, C., Park, H., and Moon, S., “What is Twitter, a Social Network or a News Media?”, Proc. of the 19th International Conference on World Wide Web, pp. 591-600, 2010.

[4] Cha, M., Haddadi, H., Benevenuto, F., and Gummadi, Krishna P., “Measuring User Influence in Twitter: The Million Follower Fallacy”, Proc. of the 4th International AAAI Conference on Weblogs and Social Media, pp. 10-17, 2010.

[5] Bernardo A., Huberman, Daniel M., Romero, and F., Wu, “Social networks that matter: Twitter under the microscope”, ArXiv, 2008.

[6] Liben-Nowell, D., Kleinberg, J., “The Link Prediction Problem for Social Networks”, Proc. fo the 12th International Conference on Information and Knowledge Management, pp. 556-559, 2004.

[7] 岡本大輝，豊田正史，喜連川優，“マイクロブログにおける対話ネットワークと投稿内容を併用したユーザ推薦に関する一考察”，情報処理学会研究報告：データベース・システム研究会報告 Vol.157, No.30, pp.1-5, 2013.