DEIM Forum 2017 D1-2

被フォロー順序に基づくユーザの役割推定手法の提案

武田　悠佑†　　佐藤　哲司††

† 筑波大学 情報学群 知識情報・図書館学類　〒305-8550 茨城県つくば市春日 1-2
†† 筑波大学 図書館情報メディア系　〒305-8550 茨城県つくば市春日 1-2
E-mail: †{ytakeda,satoh}@ce.slis.tsukuba.ac.jp

あらまし　ある話題においてユーザが情報源として果たしている役割の推定手法を提案する．本研究では，ユーザが他のユーザからフォローされる順序を被フォロー順序として，被フォロー順序にはユーザの情報源としての役割が反映されているという仮説を立てる．提案法では，フォローリストを被フォロー順序によるユーザの順序集合として，Bradley-Terry モデルよりユーザの被フォロー順序関係を抽出し，ユーザが話題において果たしている役割推定に用いる．提案法を複数の話題に対して適用した結果，役割の推定は妥当であることが示唆された．

キーワード　Twitter, 役割推定, ユーザプロファイル, フォロー順序

1. はじめに

近年，SNS（Social Networking Service）と呼ばれる，登録されたユーザ同士がコミュニケーションを行うサービスの利用者が増加している．2006 年にサービスを開始した Twitter(注1) は，2015 年 12 月時点で，約 3 億の月間アクティブユーザを持つ [1] 最も利用されている SNS の一つである．現在では，その規模の大きさからコミュニケーションの場としてのみならず，広告や広報を行う情報発信のメディアとしても盛んに活用されている．

Twitter では，流通する全ての情報がユーザによる投稿であり，フォロー，アンフォローと呼ばれる機能を用いて，投稿を閲覧したいユーザを登録，解除することで，情報を取捨選択できる．これらの機能の概要を図 1 に示す．フォローとは，継続的に投稿を閲覧したい他のユーザを登録する機能であり，フォローしたユーザの投稿は，タイムラインと呼ばれる画面上に，時系列に沿って並べて表示される．本論文では，ユーザがフォローした他のユーザを「フォロウィー」，ユーザをフォローしている他のユーザを「フォロワー」，ユーザがフォローしたフォロウィーの集合を「フォローリスト」と呼ぶ．フォローリストには，ユーザがフォロウィーをフォローした時間的な順序が保持されており，ユーザが新しく他のユーザをフォローすると，当該ユーザはフォロウィーとしてフォローリストの最後尾に追加される．一方，アンフォローとはフォロウィーとして登録していたユーザをフォローリストから削除する機能であり，アンフォローしたユーザの投稿はタイムライン上には表示されなくなる．フォロウィーをアンフォローする際も，フォローリストにおいては，当該フォロウィーが削除されるだけで，前後のフォロウィーの順序関係は保持される．ユーザはこれらの機能を用いて，自分の興味，関心に応じた投稿が自身のタイムライン上に表示されるよう，フォローリストを作成，更新している．

ユーザが Twitter を利用する目的は様々だが，多くのユーザが情報収集を利用目的としていることが明らかにされている [2]．情報収集を目的に Twitter を利用するには，ユーザは自身の求める情報を発信する他のユーザを発見する必要があるが，Twitter には膨大な量のユーザが存在するため，ユーザが適切なユーザを網羅的に探索することは困難である．この問題を解決するため，ユーザと近い興味を持つ他のユーザを発見し，フォロウィー候補として推薦する研究は数多く報告されている [3] [4] [5] [6]．

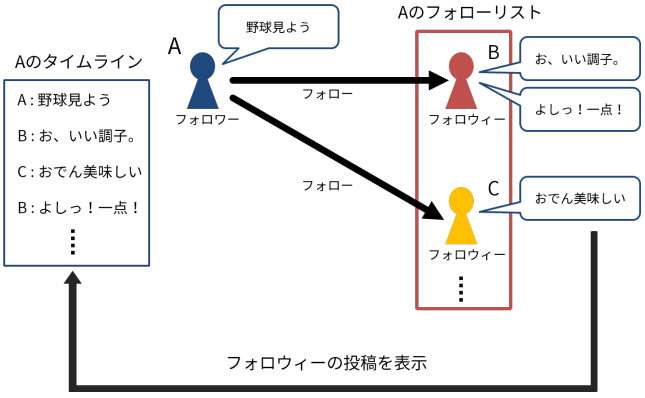

図 1　フォロー関係の概要

しかし，これまでの研究では，ユーザが特定の話題において果たしている情報源としての役割は考慮されてこなかった．同じ話題について情報を発信するユーザであっても，その話題について幅広く情報を発信するユーザと，話題の一領域について専門的な情報を発信するユーザ，話題について補足的な情報しか発信しないユーザとでは，情報源としての役割が異なると考えられる．本研究では，ユーザの情報源としての側面に着目し，ユーザが特定の話題において情報源として果たしている役割を推定することを目的とする．ユーザの役割が推定できれば，情報収集を目的とするユーザに対して，従来の手法よりも満足度の高いフォロウィー候補推薦ができると考えられるほか，広告の対象として有用であるとされる [7] 高い影響力を持つユーザの発見にも応用できると考えられる．

(注1)：https://twitter.com/

ユーザの役割を推定するにあたり，ユーザがフォロウィーをフォローする順序に着目し，ユーザがフォロウィーをフォローする順序にはユーザがフォロウィーに対して情報源として期待する役割が反映されているという仮説を立てる．つまり，ユーザがある話題について情報収集しようと複数のフォロウィーをフォローする場合，先行してフォローされるフォロウィーと，後続してフォローされるフォロウィーとでは，ユーザから情報源として期待される役割に差異があると考える．具体的には，イヤホンという話題であれば，他のフォロウィーに先行してフォローされるフォロウィーは，人気の商品や新製品について情報を発信するイヤホンの専門店のように，当該の話題について幅広く情報を発信するユーザであり，一方で，他のフォロウィーに後続してフォローされるフォロウィーは，AV（オーディオビジュアル）機器について幅広くニュースを発信するなかでイヤホンについてのニュースも扱うオンラインメディアのように，当該の話題について補足的に情報を発信するユーザである，というような違いがあると考える．このような仮説が成り立つとき，フォロウィーがフォローされる順序に着目すると，特定の話題に興味を持つ複数のユーザから，他のフォロウィーに対して同じような順序でフォローされるフォロウィーは，その話題において情報源として期待される役割が複数のユーザ間で共通するようなフォロウィーであると考えられる．情報源として期待される役割が複数のユーザ間で共通するようなフォロウィーは，実際に，その話題において情報源として何らかの役割を果たしていると考えられ，フォロウィーがフォローされる順序には，フォロウィーの情報源としての役割が反映されていると考えられる．

本研究では，このように，ユーザの被フォロー順序には情報源としての役割が反映されていると考え，フォローリストに保持されたユーザによるフォロウィーのフォロー順序を用いて，ユーザが特定の話題において情報源として果たしている役割を推定する手法を提案する．

2. 関 連 研 究

2.1 フォロウィー推薦に関する研究

Twitter ユーザを対象にフォロウィーを推薦する研究は数多く行われている．久米ら [3] は，コンテンツベースのユーザ推薦手法として，TF-IDF 法を拡張しユーザの興味に応じた重みを用いることで，より興味の近いユーザを推薦できる手法を提案している．大村 [6] は，ユーザのツイートから他の語との共起関係を基にユーザの興味語を抽出する手法を提案し，それをフォロウィー候補の推薦手法に適用することの有効性を確認している．黒柳ら [5] は，共通フォロウィーに着目し，被推薦ユーザとフォローしている著名人ユーザの重複が多いユーザをフォロウィー候補として推薦する手法を提案している．

また，ユーザ間のコミュニーケーションに着目した研究としては，北村ら [4] は，ソーシャルグラフに加えて，ユーザ間のコミュニケーションを特徴量とする手法を，坪田ら [8] が，相互フォローとなっているアカウントに対するコミュニケーションの偏りや，相互フォローを増やす速度を用いてユーザを特徴付ける手法をそれぞれ提案している．

本研究では，ある話題においてユーザが情報源として果たしている役割の推定を目的としている．フォロウィー推薦における応用では，推定された役割とともに推薦ユーザを提示することで，被推薦ユーザの情報要求をより満たす推薦が可能になると期待される．

2.2 ユーザの影響力推定に関する研究

SNS ユーザを対象としたユーザの影響力についての研究も数多く行われている．Cha ら [9] は，被フォロー数，被リツイート数，被リプライ数を用いてユーザの影響力の分析を行っている．Weng ら [10] は，PageRank を拡張した手法を用いてユーザの影響力を推定しており，その有効性を確認している．一方で，Fredrik ら [11] は，投稿に対してコメントを行うユーザの共起に着目し，Association rule learning によって，ユーザの共起を相関ルールとして学習し，学習した相関ルールに出現する回数を用いてユーザの影響力を推定する手法を提案している．Fredrik らは，精度や計算速度の観点から評価を行っているほか，新たな影響力の推定手法を提案すること自体を目的だと述べており，既存の手法とは異なるアプローチで影響力の推定への応用を考える本研究と立ち位置に類似点がある．

これらの研究に対して本研究は，ユーザの影響力を直接的に推定するのではなく，ユーザが特定の話題において果たしている役割を推定することで，間接的に影響力の高いユーザの抽出を試みる点に特徴がある．

2.3 ネットワークにおけるノードの役割に関する研究

ネットワーク上のノードをネットワーク構造に基づいてクラスタリングする研究は多数存在する．伏見ら [12] は，ネットワーク上のノードを PageRank スコアの収束を特徴量としてクラスタリングし，機能コミュニティとして抽出する手法を提案している．この研究に対して本研究は，静的なネットワーク構造ではなく，時系列データを用いてユーザの役割を推定する点に特徴がある．

また，Leon ら [13] は，社内ネットワークでの知識共有過程において，個人が果たす役割として，Knowledge diffuser, Knowledge repository, Knowledge broker, Knowledge gatekeeper という四つを挙げ，将来の知識の流れを予測するためにネットワーク構造を用いる手法を提案している．これらの研究に対して本研究は，対象とするユーザ集合が限定されたメンバーからなる社内ネットワークのノード集合ではなく，Twitter という不特定多数が参加する SNS のユーザ集合であるという点や，役割の推定自体を目的とする点に特徴がある．

3. ユーザの役割推定手法

3.1 提案手法の概要

図 2 に提案法の概要を示す．第一段階では，特定の話題に興味を持つユーザの集合を入力として，多くのユーザのフォローリストに登録されているフォロウィーを頻出フォロウィーとして抽出する．第二段階では，頻出フォロウィーをフォローリストにおける被フォロー順序についてランキングする．被フォロー順序についてのランキングには Bradley-Terry モデル

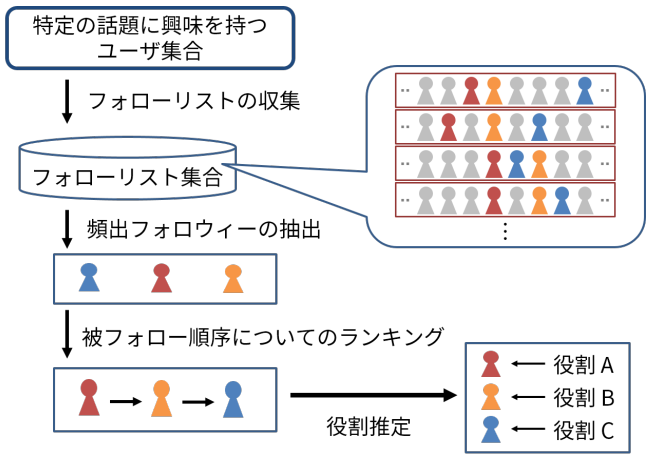

図 2　提案手法の概要

を用いる．第三段階では，ランキングを用いて，当該話題における頻出フォロウィーの役割を推定する．提案法は，フォローリストをフォロー順序によるユーザの順序集合とみなし，複数のフォローリストから，頻出フォロウィーについての一つの順序関係を導出する点に特徴がある．役割の推定では，頻出フォロウィーの，被フォロー順序に基づくランキング順位とフォロワー数によるランキング順位との差に基づいてユーザの役割を限定し三つに分類する．

3.2　頻出フォロウィーの抽出

入力となる特定の話題に興味を持つユーザ集合を U としたとき，各ユーザ $u \in U$ のフォローリスト $list(u)$，フォローリスト集合 $FollowList$，フォロウィー集合 $Followee$，頻出フォロウィー集合 $V^{(q)}$ を以下のように定義し，それぞれ求める．$list(u)$ は Twitter REST API を用いて取得する．

$$
\begin{aligned}
list(u) &= \{followee \,|\, u \text{ follows } followee\} \\
FollowList &= \{list(u) \,|\, u \in U\} \\
Followee &= \cup_{u \in U}\, list(u)
\end{aligned}
$$

$$
V^{(q)} = \{followee \in Followee \,|\, \frac{|\{list \,|\, followee \in list\}|}{|FollowList|} > q\}
$$

ここで，抽出される頻出フォロウィー $v \in V^{(q)}$ が役割推定の対象ユーザとなり，q は $Followee$ 抽出される $V^{(q)}$ を調整するパラメータである．定義式から q の値を小さくすれば，$V^{(q)}$ は大きくなる一方で，話題について関連性の低いユーザが抽出されるようになると考えられる．反対に，q の値を大きくすれば $V^{(q)}$ は小さくなる一方で，話題により関連性の高いユーザを抽出できると考えられる．

3.3　被フォロー順序についてのランキング

3.2 節で得られる $V^{(q)}$ について，Bradley-Terry モデル [14] を用いて，フォローリスト毎に存在するフォロウィー同士の被フォロー順序関係から，被フォロー順序についての一つのランキングを作成する．Bradley-Terry モデルとは，複数の順序関係から一つの順序関係を導出するモデルであり，複数のランキングを統合して，一つのランキングを作ることができることから，スポーツチームの勝敗予測などに用いられる手法である．

Bradley-Terry モデルでは，ある集合における各要素について，各要素間の順序関係を確率的に決定する強さに相当する値が存在すると仮定し，その値を複数の順序関係から推定する．推定した値に基づき各要素をランキングすることで，複数の順序関係から一つの順序関係を導出する．提案法では，ユーザには他のユーザに対して，あるフォローリスト中においてどちらが先にフォロウィーとしてフォローされるかを式 (1) のように確率的に決定する値 x が存在すると仮定する．ただし，ここで i beats j とは，同一のフォローリスト中においてユーザ i がユーザ j よりも前に出現する，つまりあるユーザによって i が j に先行してフォローされることを表すものとする．

$$
P(i \,\text{beats}\, j) = \frac{x_i}{x_i + x_j} \tag{1}
$$

フォローリストを，被フォロー順序によって順序付けられるフォロウィー間の順序関係とみなして，Bradley-Terry モデルを適用し，頻出フォロウィーについての x の系列 $\mathbf{x}$ を推定する．$\mathbf{x}$ の推定値に基づき，$V^{(q)}$ における被フォロー順序関係を表したランキング $Rank$ を作成する．

$\mathbf{x}$ は以下の手順で推定する．フォロウィー i と j が同時に出現するフォローリスト数を N_{ij}，その内 i beats j となっているフォローリスト数を n_{ij} とおくと，フォローリストに保持されたフォロー順序は全順序であるため，式 (2) のように書ける．

$$
N_{ij} = n_{ij} + n_{ji} \tag{2}
$$

ここで，n_{ij} が二項分布 $B(N_{ij}, P(i \,\text{beats}\, j))$ に従うと仮定し，$m = |V^{(q)}|$ とおき，$n_{ii} = 0$ であるとすると，$\mathbf{x}$ の尤度関数 L は式 (1) より，

$$
\begin{aligned}
L(\mathbf{x}) &= \prod_{i=1}^{m}\prod_{j=1}^{m} P(i \,\text{beats}\, j)^{n_{ij}} \\
&= \prod_{i=1}^{m}\prod_{j=1}^{m} \left(\frac{x_i}{x_i + x_j}\right)^{n_{ij}} \\
&= \prod_{i=1}^{m}\prod_{j>i}^{m} (x_i + x_j)^{-n_{ij}} \cdot (x_i + x_j)^{-n_{ji}} \cdot {x_i}^{n_{ij}} \cdot {x_j}^{n_{ji}} \\
&= \prod_{i=1}^{m}\prod_{j>i}^{m} (x_i + x_j)^{-N_{ij}} \cdot \prod_{i=1}^{m} {x_i}^{W_i} \qquad \left(\text{s.t. } W_i = \sum_{j}^{m} n_{ij}\right)
\end{aligned}
$$

と書け，この尤度関数の対数を取った対数尤度関数 l は，式 (3) となる．ここで，$a > 0$ なる a については，$l(a\mathbf{x}) = l(\mathbf{x})$ となることに注意し，$\mathbf{x}$ について，式 (4) の制約を設ける．

$$
l(\mathbf{x}) = \sum_{i=1}^{m} W_i \log x_i - \sum\sum_{i<j} N_{ij} \log (x_i + x_j) \tag{3}
$$

$$
\sum_{i=1}^{m} x_i = 1 \tag{4}
$$

式 (4) の下で対数尤度関数 l に対してラグランジュの未定乗数法を用いると，ラグランジュ関数 $\tilde{l}$ はラグランジュ乗数を λ として，

$$
\tilde{l} = \sum_{i=1}^{m} W_i \log x_i - \sum\sum_{i<j} N_{ij} \log (x_i + x_j) - \lambda \left(\sum_{i=1}^{m} x_i - 1\right)
$$

と書ける．これを x_i, λ で偏微分すると，式 (6) となり，式 (7)，式 (8) が得られる．

$$\frac{\partial}{\partial x_i}\left\{l-\lambda\left(\sum_{i=1}^{m}x_i-1\right)\right\}=0 \tag{5}$$

$$\frac{\partial}{\partial \lambda}\left\{l-\lambda\left(\sum_{i=1}^{m}x_i-1\right)\right\}=0 \tag{6}$$

$$\frac{W_i}{x_i}-\sum_{j\neq i}^{m}\frac{N_{ij}}{x_i+x_j}-\lambda=0 \tag{7}$$

$$\sum_{i=1}^{m}x_i-1=0 \tag{8}$$

式 (7) より，式 (9) が得られ，これを全ての i について足し合わせると，式 (10) となる．

$$W_i=\lambda x_i+\sum_{j\neq i}^{m}N_{ij}\frac{x_i}{x_i+x_j} \tag{9}$$

$$\begin{aligned}\sum_{i=1}^{m}W_i &= \lambda\sum_{i=1}^{m}x_i+\sum_{i=1}^{m}\sum_{j\neq i}^{m}N_{ij}\frac{x_i}{x_i+x_j}\\ &= \lambda\sum_{i=1}^{m}x_i+\sum_{i=1}^{m}\sum_{j>i}^{m}N_{ij}\left(\frac{x_i}{x_i+x_j}+\frac{x_j}{x_i+x_j}\right)\\ &= \lambda+\sum_{i=1}^{m}\sum_{j>i}^{m}N_{ij}\end{aligned} \tag{10}$$

ここで式 (10) に対して，

$$\sum_{i=1}^{m}W_i=\sum_{i}^{m}\sum_{j>i}^{m}N_{ij} \tag{11}$$

であることを用いると，$\lambda=0$ であることがわかる．したがって，

$$\sum_{j\neq i}N_{ij}\frac{x_i}{x_i+x_j}=W_i \tag{12}$$

$$\sum_{i=1}^{m}x_i=1 \tag{13}$$

の連立方程式が解くべき尤度方程式となる．尤度方程式を解くために，$\mathbf{x}^{(0)}=\{{x_1}^{(0)},{x_2}^{(0)},\cdots,{x_m}^{(0)}\}$ を適当な初期値として，i 毎に式 (12) を変形した式 (14) の右辺の近似値を，x_i の推定値 ${x_i}^{(1)}$ として求める．

$$\tilde{x_i}=\frac{W_i}{\sum_{j\neq i}\frac{N_{ij}}{x_i+x_j}} \tag{14}$$

求めた推定値が式 (15) を満たすまで，推定値を式 (14) に対して適用し，順次推定値を更新する．

$$\sum_{i=1}^{m}\left({x_i}^{(k+1)}-{x_i}^{(k)}\right)^2<(20m)^{-1} \tag{15}$$

式 (15) を満たした時点での推定値 $\tilde{\mathbf{x}}$ を用いて，フォロウィーを降順に並べることで，頻出フォロウィーの被フォロー順序についてのランキングが求められる．

表 1　データセット概要

	機械学習	イヤホン
代表ユーザ	@shima__shima	@kindo3
フォロワー数	3,699	3,414
データ収集日	2016 年 8 月 14 日	2016 年 8 月 17 日

3.4　ランキングに基づく役割推定

ユーザの役割を推定するに際して，フォロワー数の多さと，被フォロー順序の関係に着目する．一般に，フォロワー数が多いユーザほどよく知られたユーザであると考えられ，よく知られたユーザほどフォロウィーの候補となる確率が高いと考えられる．つまり，フォロワー数が多いユーザほどより早期にフォローされる傾向にあると考えられる．ここで，フォロワー数がそれほど多くないにも関わらず早期にフォローされる傾向にあるユーザや，逆に，フォロワー数が多いにも関わらず後からフォローされる傾向にあるユーザは，何らかの特徴を有するユーザではないかと考えられる．提案法では，3.3 節で作成するランキングと，フォロワー数によるランキングにおける順位の差に着目し，差が大きいユーザは情報源として特徴的な役割を果たしているユーザであるとして，以下の式によりユーザの役割を推定する．$FollowerRank$ とは，$V^{(q)}$ をフォロワー数により降順でランキングしたもの，$Rank(i)$ とは，ユーザ i の $Rank$ における順位，$FollowerRank(i)$ とは，i の $FollowerRank$ における順位，σ とは，$(Gap(1), Gap(2), \cdots, Gap(m))$ の標準偏差であり，$Gap(i)=Rank(i)-FollowerRank(i)$ である．

$$Role(i)=\begin{cases}A & (if\ Gap(i)>\sigma)\\ B & (if\ Gap(i)<-\sigma)\\ C & (otherwise)\end{cases}$$

4.　実験と評価

実験では，複数の話題を対象に，提案法を適用し，作成されるランキングと抽出される役割の性質を評価する．ランキングの評価では，提案法における，フォロワー数が多いユーザほど早期にフォローされる傾向にあるとの仮定を検証するほか，ユーザの他の特徴量との相関も分析する．役割の評価では，各役割に割り当てられたユーザがどのような特徴を持つか評価し，それぞれの役割について検討する．

4.1　データセット

本論文では，機械学習とイヤホンという分野の異なる二つの話題を対象に，実験を行った．各話題において入力となるユーザ集合は，人手により選択した，話題を代表するユーザのフォロワーとした．表 1 にデータセットの概要を示す．

4.2　評 価 方 法

ランキングについては，ランキングにおけるユーザの順位とユーザの様々な特徴量との順位相関を求める．順位相関係数にはスピアマンの順位相関係数を使用する．役割については，役割毎に，割り当てられたユーザの特徴量の平均値を算出し，役割間の関係からそれぞれの役割を特徴付ける．

ユーザの特徴量としては，総投稿数，フォロウィー数，フォロワー数，リスト数，アカウント作成日，頻出フォロウィー中のフォロウィー数，頻出フォロウィーからの被フォロウィー数，一ヶ月あたりのツイート数，リプライ数，リツイート数，URLを含むツイート数，ハッシュタグを含むツイート数，ツイートのトピック分布のエントロピーを用いる．本論文では，Twitterにおけるユーザの投稿をツイート，リプライ，リツイートの三種類に分類する．リプライとは，本文中に送り先となるユーザのユーザ ID（スクリーンネーム）を含めることにより特定のユーザに向けてメッセージを投稿する機能，またはその機能によってなされた投稿であり，リツイートとは，過去になされた投稿を投稿者の自他を問わず転載する機能，またはその機能によってなされた投稿であるとする．リプライ，リツイートのいずれでもない通常の投稿はツイートと呼ぶこととする．また，リストとは，ユーザがユーザ集合をフォローリストとは別に作成できる機能であり，その機能によって作られたユーザ集合もリストと呼ばれる．リストにはフォロウィーでないユーザや自らのアカウントも含めて任意のユーザを登録することができ，リストを作成すると当該リストに登録されたユーザの投稿のみが表示される画面を閲覧できるようになる．ハッシュタグとは，ユーザが自分のツイートに対して自由にタグをつけることができる機能，またはその機能によって作成されたタグのことである．ハッシュタグは#記号を文字列の先頭につけることで作成でき，主にツイートの話題を明示的に識別する目的で使用される．

4.3 ユーザ特徴量の抽出

本節では評価に用いるユーザの特徴量とその抽出法について説明する．総投稿数とは，リプライ，リツイートを含めたユーザの投稿の総数，フォロウィー数とは，ユーザがフォローしているフォロウィー数，フォロワー数とは，ユーザをフォローしているフォロワー数，リスト数とは，ユーザが登録されているリスト数，アカウント作成日とは，ユーザのアカウント作成日であるとする．これら 5 つの特徴量については，Twitter REST API を用いて直接的に取得する．

頻出フォロウィー中のフォロウィー数とはユーザのフォロウィー内の頻出フォロウィー数，頻出フォロウィーからの被フォロウィー数とはユーザのフォロワー内の頻出フォロウィー数であるとする．これら 2 つの特徴量については，頻出フォロウィーの二人組の全組み合わせについて，Twitter REST API を用いてフォロー関係の有無を取得し，値を算出する．

一ヶ月あたりのツイート数，リプライ数，リツイート数，URLを含むツイート数，ハッシュタグを含むツイート数は，Twitter REST API を用いてユーザが一ヶ月間に行った全投稿を取得し，それぞれ値を抽出する．本論文では，各頻出フォロウィーについて，2016 年 11 月 1 日から 2016 年 11 月 30 日までの 30 日間を対象に投稿を取得し実験に用いた．ツイートのトピック分布のエントロピーの抽出法については次節で述べる．

4.4 トピック分布のエントロピー

ツイートのトピック分布のエントロピーの抽出は，ツイートのトピック分布の推定と，トピック分布のエントロピー算出の二段階からなる．ツイートのトピック分布の推定では，ユーザによる複数のツイートを結合したものに対して，トピックモデルの一種である HDP-LDA モデル [15] を適用して，トピック分布を推定する．トピック分布のエントロピー算出では，得られたトピック分布の平均情報量をツイートのトピック分布のエントロピーとして算出する．ここで算出されるツイートのトピック分布のエントロピーとは，ユーザのツイートが属するトピックのあいまいさを表しており，ユーザのツイートのトピック分布の偏りが大きいほどその値は小さくなる．つまり，ツイートのトピック分布のエントロピーの値の大きさは，ユーザが発信する情報の幅広さに比例すると考えられる．

表 2　フォローリスト概要

	機械学習	イヤホン
収集フォローリスト数	3156	3002
フォローリストの平均長	606.4	617.5
総フォロウィー数	869,394	857,282

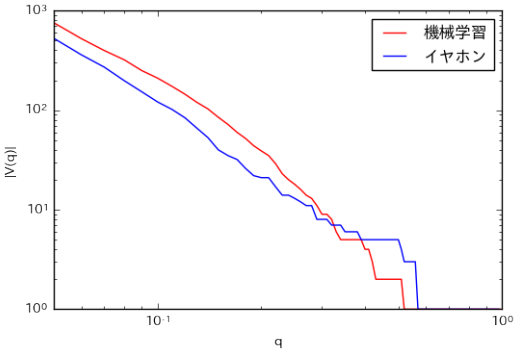

図 3　$|V^{(q)}|$ の推移

本論文では，ユーザによるツイートを 1000 件収集し，それぞれ日本語形態素解析器である MeCab[(注2)] を用いて名詞のみを抽出し結合したものを一文書とみなし，トピック分布を推定した．HDP-LDA モデルは，日本語版 Wikipedia[(注3)] の記事から MeCab によって名詞以外を取り除き，50%以上の記事に出現する名詞をストップワードとみなして更に取り除いたものをコーパスとして構築した．HDP-LDA モデルの構築には，gensim[(注4)] を用いた．

4.5 実験結果

本節では，実験結果について示す．実験では実装の都合上，提案法の第一段階となるフォローリストの収集では，フォロウィー数が 5000 以上のユーザをスパムユーザとして，フォローリストの収集対象から除外した．投稿を非公開にしているユーザについても，フォローリストの収集はできないため，フォローリストの収集対象から除外した．各話題において収集したフォローリストの概要を表 2 に示す．

図 3 に，それぞれの話題における，q の値と抽出される頻出フォロウィー集合 $V^{(q)}$ の大きさの関係を示す．適切な q の値

(注2)：http://taku910.github.io/mecab/
(注3)：https://ja.wikipedia.org/wiki/
(注4)：https://radimrehurek.com/gensim/

表 3　*Rank* 作成結果

順位	機械学習	イヤホン
1	@masason	@e_earphone
2	@yukihiro_matz	@eear_ryouta
3	@hyuki	@FUJIYAAVIC
4	@mamoruk	@eear_takuya
5	@shima__shima	@iriver_Lyumo
6	@hillbig	@avwatch
7	@ibisml	@fitear
8	@hamadakoichi	@OYAIDE_NEO
9	@preferred_jp	@phileweb
10	@AntiBayesian	@kindo3
11	@sla	@nomurakenji
12	@unnonouno	@KumitateLab
13	@sylvan5	@KumitateK
14	@TJO_datasci	
15	@neubig	
16	@issei_sato	
17	@iwiwi	
18	@beam2d	

q_f として，各話題において，q の値を 0.05 から 1.00 まで 0.01 ずつ動かし，各時点における $|V^{(q)}|$ の値を算出し，各時点における $|V^{(q)}|$ の値と，その一時点前の $|V^{(q)}|$ の値の差が閾値 s 未満になった時点の q の値を求める．本論文では，$s = 3$ として実験したところ，機械学習の話題では $q_f = 0.25$，イヤホンの話題では $q_f = 0.24$ となった．両話題の q_f の値を揃えるため，$q_f = 0.25$ として，以後の実験を行った．

各話題において作成された被フォロー順序についてのランキング *Rank* を表 3 に示す．表 3 より，機械学習の話題では 18 ユーザが，イヤホンの話題では 13 ユーザが頻出フォロウィーとして抽出されたことが確認できる．次に，各話題における頻出フォロウィーの，*Rank* における順位とユーザの各特徴量との順位相関係数を表 4 に示す．表 4 中で，左の端の列はユーザの特徴量の種類を，中央の列は，機械学習の話題における頻出フォロウィーの，*Rank* における順位と特徴量との順位相関係数を，右端の列は，イヤホンの話題における頻出フォロウィーの，*Rank* における順位と特徴量との順位相関係数を表している．表 4 より，どちらの話題においても，フォロワー数と被フォロー順序によるランキングの順位との間に，負の相関関係が認められた．

各話題における役割の推定結果を表 5 に示す．機械学習における σ の値は 4.67，イヤホンにおける σ の値は 2.77 であった．各話題における役割毎の特徴量の平均値を表 6 に示す．ただし，アカウント作成日については平均値を算出しなかった．

5. 考　　察

5.1　ランキングの性質評価

表 4 より，被フォロー順序についてのランキング順位とフォロワー数とは負の相関関係にあることを確認できる．この結果より，フォロワー数が多いユーザほど早期にフォローされる傾向にあるとの仮定は誤りではないと考えられる．一方，アカウント作成日に着目すると，両話題において，被フォロー順序についてのランキング順位との相関は認められないことが確認できる．これは，ユーザはフォローリストをフォローだけでなくアンフォローによっても随時更新しているためではないかと考えられる．

表 4　*Rank* 順位と各特徴量の順位相関係数

	機械学習	イヤホン
総投稿数	−0.10	−0.47
フォロウィー数	0.47	−0.56*
フォロワー数	−0.60**	−0.73**
リスト数	−0.64**	−0.69**
アカウント作成日	0.29	0.43
ツイートのトピック分布のエントロピー	0.25	0.19
頻出フォロウィーへのフォロー数	0.49*	−0.71**
頻出フォロウィーからの被フォロー数	0.36	−0.13
ツイート数/month	0.03	−0.60*
リプライ数/month	−0.07	−0.11
リツイート数/month	0.23	−0.75**
URL を含むツイート数/month	−0.31	−0.54
ハッシュタグを含むツイート数/month	0.00	−0.82**

*: $p < .05$, **: $p < .01$

表 5　役割推定結果

役割	機械学習	イヤホン
A	@shima__shima	@eear_ryouta
	@ibisml	@eear_takuya
	@preferred_jp	
B	@AntiBayesian	@avwatch
	@TJO_datasci	@OYAIDE_NEO
	@iwiwi	@phileweb
C	@masason	@e_earphone
	@yukihiro_matz	@FUJIYAAVIC
	@hyuki	@iriver_Lyumo
	@mamoruk	@fitear
	@hillbig	@kindo3
	@hamadakoichi	@nomurakenji
	@sla	@KumitateLab
	@unnonouno	@KumitateK
	@sylvan5	
	@neubig	
	@issei_sato	
	@beam2d	

5.2　各役割の性質評価

表 6 より，両話題において，役割 *B* におけるフォロワー数は役割 *A* よりも大きいことが確認できる．一方で，頻出フォロウィーからの被フォロー数は，単純なフォロワー数は役割 *B*，役割 *C* が上回っていたのに対して，役割 *A* における値が役割 *B*，役割 *C* における値よりも大きくなっていることが両話題で確認できる．頻出フォロウィーは，当該話題に興味を持つ多くのユーザからフォローされている，その話題における有用な情報源だと考えられるユーザである．このような頻出フォロウィーの多くからフォローされている役割 *A* のユーザは，当該話題における権威的なユーザであることが考えられる．また，

表 6　役割毎の特徴量

	機械学習			イヤホン		
	役割 A	役割 B	役割 C	役割 A	役割 B	役割 C
総投稿数	10,528	21,807	28,248	26,056	25,851	30,942
フォロウィー数	411	555	915	2,006	3,973	4,978
フォロワー数	3,235	7,586	221,121	6,861	19,382	7,671
リスト数	234	497	5,698	201	1,348	264
ツイートのトピック分布のエントロピー	1.57	1.75	1.54	0.76	0.62	1.06
頻出フォロウィーへのフォロー数	6.33	7.00	9.17	12.00	5.33	9.25
頻出フォロウィーからの被フォロー数	9.67	8.00	8.08	9.50	7.67	9.00
ツイート数/month	101.00	223.00	300.50	567.50	422.00	516.00
リプライ数/month	60.00	26.33	91.92	157.50	1.33	120.88
リツイート数/month	2.67	120.33	104.67	96.50	69.33	177.00
URL を含むツイート数/month	27.00	21.00	45.33	43.50	347.67	140.12
ハッシュタグを含むツイート数/month	10.33	3.00	8.17	50.50	8.67	63.50

一ヶ月あたりのツイート数に対するリツイート数，URL を含むツイート数の割合に着目すると，イヤホンの話題でこそ役割 B よりもリツイート数の割合が大きくなっているものの，役割 A が他の役割と比較し相対的に低い値を示すことが確認できる．この結果から，役割 A のユーザは，より一次的な情報を発信する傾向にあることが伺える．役割 C については，両話題において，フォロウィー数が他の二つの役割と比較して大きいことが確認できる．この結果は，役割 C のユーザは，他の役割のユーザに比べて，情報収集を目的とする傾向が強いことを示していると考えられる．

役割 A を割り当てられたユーザについて人手により素性を調査すると，@shima__shima は機械学習について資料をまとめて公開している機械学習を専門とする研究者のアカウント，@ibisml は機械学習を対象とした研究会の公式アカウント，@preferred_jp は機械学習や自然言語処理を用いた製品を開発する会社の公式アカウント，@eear_ryouta，@eear_takuya は共にイヤホン専門店の店員のアカウントであることが判明した．同様に，役割 B に割り当てられたユーザについて人手により素性を調査すると，@AntiBayesian，@TJO_datasci はデータサイエンティストのアカウント，@iwiwi は離散アルゴリズムを専門とする研究者のアカウント，@avwatch，@phileweb はオーディオビジュアルに関係する情報を扱うウェブメディアの公式アカウント，@OYAIDE_NEO はオーディオアクセサリーの製造，販売店の公式アカウントであることが判明した．役割 A を割り当てられたユーザはいずれも，他ユーザへの積極的な情報発信を行っている，当該話題における専門家的なユーザであった．一方，役割 B を割り当てられたユーザはいずれも当該の話題についての情報も発信するものの，専門とする話題は当該話題の周辺分野であるようなユーザであった．このような結果から，定性的にではあるが，提案手法により，特定の話題において特徴的な役割を果たしているユーザを抽出できることが示唆されたと考える．

6. おわりに

本論文では，フォローリストにおける順序性に着目し，Twitter ユーザがある話題において果たしている役割の推定手法を提案した．ユーザの役割推定に際して本研究では，ユーザがフォロウィーをフォローする順序には，フォロウィーがある話題において果たしている役割が反映されており，フォローの時間的な順序が保存されたフォローリストにはフォロウィーの役割が反映されているという仮説を立てる．提案手法は，フォローリストをユーザの順序集合とみなすことにより，フォローリスト集合を Bradley-Terry モデルの学習データとして使用し，ユーザ間の被フォロー順序をランキングする点に特徴がある．第一段階では，フォローリスト集合から頻出フォロウィーを抽出し，第二段階では，頻出フォロウィーを Bradley-Terry モデルによって被フォロー順序でランキングする．最終段階となる第三段階では，被フォロー順序のランキングとフォロワー数のランキングにおける順位の差を用いて，ユーザの役割を A，B，C の三つに分類する．

機械学習とイヤホンという分野の異なる二つの話題に対して実験を行った結果，フォロワー数が多いユーザほど早期にフォローされる傾向にあること，より以前からアカウントが存在するユーザほど早期にフォローされるという傾向は顕著ではないことが明らかになった．また，役割を A だと推定されるユーザは，他ユーザへの積極的な情報発信を目的とする当該話題における専門家的なユーザであり，頻出フォロウィーから多くのフォローを集める傾向にあることが示唆された．一方で，役割を B だと推定されるユーザは，当該話題についての情報も補足的に発信するものの，当該話題の周辺分野を専門とするユーザであること，役割を C だと推定されるユーザは，情報収集を目的とする傾向が強いことが示唆された．このような結果より，提案法により，特定の話題において特徴的な役割を果たしているユーザを抽出できることが示唆された．

今後の展望としては，提案法による役割の推定が妥当であることを示すより進んだ評価手法や，ユーザのフォロワーとしての側面も考慮したモデルの高度化，改良が考えられる．

謝　辞

本研究は JSPS 科研費 JP16H02904 の助成を受けたものです．

文　　献

[1] 安藤健二，“Twitter が国内ユーザー数を初公表 「増加率は世界一」”．The Huffington Post, http://www.huffingtonpost.jp/2016/02/18/twitter-japan_n_9260630.html, 参照 Dec. 11,2016.

[2] “オプト、twitter（ツイッター）の利用実態に関する調査を実施”．オプト，http://www.opt.ne.jp/news/pr/detail/id=2341, 参照 Dec. 11,2016.

[3] 久米雄介，打矢隆弘，内匠逸，“興味領域を考慮した Twitter ユーザ推薦手法の提案と評価，” 情報処理学会研究報告，vol.2015-ICS-179，no.1，pp.1–8，Mar. 2015.

[4] 北村太一，小川祐樹，諏訪博彦，太田敏澄，“ユーザ間関与に基づく Twitter フォローユーザ推薦，” 日本社会情報学会全国大会研究発表論文集，vol.26，pp.229–232，2011. DOI:10.14836/jasi.26.0.229.0

[5] 黒柳智士，山田泰宏，鈴木浩，服部哲，速水治夫，“著名人情報に基づいた Twitter フォローユーザ推薦システム，” 情報処理学会研究報告，vol.2013-CDS-6，no.25，pp.1–4，Jan. 2013.

[6] 大村涼，赤石美奈，佐藤健，“語彙連鎖構造を用いた twitter ユーザー推薦手法の提案，” 情報処理学会第 75 回全国大会講演論文集，vol.2013，no.1，pp.609–610，2013.

[7] M. Trusov, A.V. Bodapati, and R.E. Bucklin, “Determining influential users in internet social networks,” Journal of Marketing Research, vol.47, no.4, pp.643–658, 2010.

[8] 坪田啓司，小林亜樹，“Twitter にて会話しやすいユーザを推薦する手法の評価，” 情報処理学会研究報告，vol.2013-DBS-158，no.15，pp.1–8，Nov. 2013.

[9] M. Cha, H. Haddadi, F. Benevenuto, and K.P. Gummadi, “Measuring user influence in twitter: The million follower fallacy,” Proceedings of international AAAI Conference on Weblogs and Social Media, pp.10–18, 2010.

[10] J. Weng, E.-P. Lim, J. Jiang, and Q. He, “Twitterrank: finding topic-sensitive influential twitterers,” Proceedings of the third ACM international conference on Web search and data mining, pp.261–270, 2010.

[11] F. Erlandsson, P. Bródka, A. Borg, and H. Johnson, “Finding influential users in social media using association rule learning,” Entropy, vol.18, no.5, p.164, 2016.

[12] 伏見卓恭，斉藤和巳，風間一洋，“ネットワーク機能コミュニティ抽出法，” 日本データベース学会論文誌，vol.10，no.3，pp.13–18，2012.

[13] R.-D. Leon, R. Rodríguez-Rodríguez, P. Gómez-Gasquet, and J. Mula, “Social network analysis: A tool for evaluating and predicting future knowledge flows from an insurance organization,” Technological Forecasting and Social Change, vol.114, pp.103–118, Jan. 2017.

[14] R.A. Bradley and M.E. Terry, “Rank analysis of incomplete block designs: I. the method of paired comparisons,” Biometrika, vol.39, no.3/4, pp.324–345, 1952.

[15] Y.W. Teh, M.I. Jordan, M.J. Beal, and D.M. Blei, “Hierarchical dirichlet processes,” Journal of the American Statistical Association, vol.101, no.476, pp.1566–1581, 2006.