フォロイーのツイートパターン分析に基づいたユーザ推薦システム

鴨下　海人†　　北山　大輔†

† 工学院大学情報学部コンピュータ科学科　　〒 163–8677 東京都新宿区西新宿 1–24–2
E-mail: †j111032@ns.kogakuin.ac.jp, ††kitayama@cc.kogakuin.ac.jp

あらまし　近年，マイクロブログのユーザが増加傾向にある．その一つである Twitter では，他のユーザをフォローすることで，そのユーザのツイートを見ることができるようになる．しかしながら，Twitter のユーザは膨大であるため，フォローするべきユーザを見つけるのは困難である．このことから，Twitter 上でフォローするべきユーザを推薦する研究は数多くあるが，フォロー数が多すぎると，タイムライン上の情報量が増えるために，かえって閲覧しにくくなってしまう．そこで，本論文では，より精度の高い推薦を実現するべく，フォロイーのツイートの分類を行い，推薦に加味することを提案する．また，本手法の有効性を確かめるべく，普段から Twitter を利用しているユーザを対象に評価実験を行った．その実験の結果と考察についても報告する．

キーワード　Twitter, ユーザ推薦, マイクロブログ, ツイート分析

1. はじめに

近年，Twitter というマイクロブログが広く利用されるようになってきた．Twitter では,「ツイート」という，140 文字以下の記事を Web に投稿することで，情報を発信したり，他のユーザのツイートを閲覧することで，情報を受信したりすることができる．また，ツイートは，特定のユーザに向けて発信する「リプライ」という形態をとることもできる．これは，対象のユーザの ID（ユーザごとにユニークな英数字．記号はアンダーバーのみ使用可能）をツイート中に含むことによって，対象のユーザに通知されるツイートである．さらに,「フォロー」という機能も存在し，ユーザは他のユーザをフォローすることで，そのユーザのツイートを簡単に見ることができるようになる．いわば，ブックマークのような機能である．このフォローという機能が，Twitter を特徴づけている重要な要素であり，ユーザはフォローや「リムーブ（フォローを解除すること）」を繰り返すことで，欲しい情報や見たい情報に合わせて，好きなように，自身の「タイムライン（フォローしているユーザのツイートが表示される，ホームページのようなもの）」をカスタマイズすることができる．

ユーザがタイムラインを自由にカスタマイズするということは，あるユーザ A がフォローしているユーザ B（これをフォロイーと呼ぶ）には，ユーザ A がフォローしたときの意図があることになる．また，ユーザのフォロイーは複数存在することから，それぞれのフォロイーにそれぞれの意図があると考えられる．我々は，そこから，ユーザ A の Twitter の使い方を推定することができると考えた．たとえば，スポーツについての最新情報を知るために Twitter を利用しているユーザは，スポーツについて最新情報をツイートするユーザをフォローすると考えられ，また，誰かとネットを介して会話するための，チャット感覚で Twitter を使うユーザは，会話を好むユーザをフォローすると考えられる．

そこで，本研究では，ユーザの Twitter の使い方に着目し，既存のユーザ推薦とは異なる視点を用いたユーザ推薦を提案する．2 章で提案する手法の概要と，ユーザ推薦に関連する研究の紹介を行う．3 章では，提案手法の具体的なアルゴリズムと，その実現方法を述べる．4 章で，Twitter ユーザによる評価実験と，その考察を行い，最後に 5 章で，まとめと今後の展望について触れていく．

2. 研究のアプローチ

2.1 概　　要

本研究においては，ユーザ自身が何についてツイートしているのかは，ユーザの Twitter の使い方に関係がないと考えている．先程の例であれば，主にスポーツの最新情報が知りたいがために Twitter を利用しているユーザが，必ずしもスポーツの最新情報をツイートするとは限らず，また，チャットについても，実際に自分はチャットに参加せず，やり取りを見て楽しみたいという人もいる．そこで我々は，ユーザの Twitter の使い方の違いは，フォロイーの違いとなって表れると考え，ユーザがフォロイーを選定する際には，ユーザの Twitter の使い方が影響していると考えた．

そこで，本研究では，ユーザのフォロイーに注目し，フォロイーのツイートの分類を行うことで，ユーザの Twitter の使い方を考慮し，それに沿ったユーザ推薦を実現する手法を提案する．ここで，提案手法では，ユーザの Twitter の使い方のみを推定の対象としていることから，すべてのユーザの中から，提案手法のみによってユーザを推薦することは難しい．そこで，既存の，何らかの基準によって推薦されるユーザのリストを用いて，そのリストをリランキングすることにより，目的である，より精度の高いユーザ推薦を実現することとする．以降，このリストに存在するユーザを推薦ユーザと呼ぶ．

提案手法の手順として，まず，手法の対象となるユーザのフォロイーのツイートを取得し，あらかじめ定義した種類に分ける．次に，ツイートの種類ごとの割合を求め，それをフォロイーのツイートパターンとして抽出する．最後に，推薦ユーザ

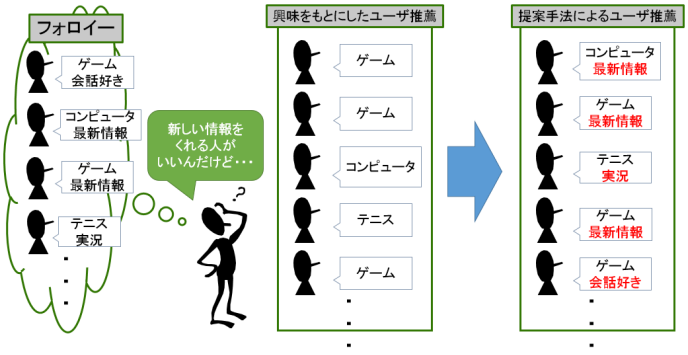

図 1　提案手法の概念図

表 1　該当ツイートの具体例

種類	該当ツイート
情報発信	インクは血よりも高い，プリンターのインクカートリッジがどれだけ高値なのかが一目で分かるグラフ - http://dummy.com インクが高いのは分かってたけど，シャネルの 5 番ってその 10 倍するんすね…
情報要求	ユーザに固有の，@以下の英数字あるじゃないですか あれって普通なんて呼ぶ？
会話	@dummy_account 自然薯とジャネンバって似てません？
その他	紙媒体の論文読んでてページめくるためにマウスに手を伸ばしてマウスホイール転がして一人で爆笑した

のツイートパターンと，フォロイーのツイートパターン間の類似度を算出し，これに基づいて，推薦ユーザをランキング形式で出力する．これにより，ユーザの Twitter の使い方に沿ったユーザを推薦することができるようになる．図 1 は提案手法の概念図である．

2.2　関 連 研 究

Twitter を対象にして，ユーザの分類を行っている研究が存在する [1] [2] [3]．たとえば，竹村らの研究 [1] では，ユーザのフォロワーに着目することで，ユーザが，広く一般のユーザが興味を示す情報を発信しているのか，それとも，一部のユーザが興味を示す情報を発信しているのかを判定し，分類している．本研究では，情報を発信するようなツイートを求めているユーザが存在すると考え，推薦に反映しているが，その内容や範囲の別については考慮していない．また，田中ら [2] の研究では，あるユーザがあるユーザをフォローする際の意図が，フォロイーごとに存在するとしている．本研究では，ユーザごとに，意図がある程度のまとまりを持つと仮定して，それに沿ったユーザ推薦が行えるのではないかと考えている．

ユーザ推薦の研究として，康ら [4] は，ユーザは自分と同じ属性を持つユーザをフォローするとして，属性伝搬モデルを構築することによって，ユーザ推薦を行っている．しかし，本研究では，ユーザがフォローするユーザは，必ずしもユーザ自身の属性によらないと考えており，フォロイーに着目してユーザ推薦を行う．また，フォローのネットワークを用いたグラフを利用するのではなく，ツイートの分類によって推薦を実現している．この他にも，ユーザ推薦手法は Twitter 上のフォローもしくは対話のネットワークに基づくものが多い．[5] [6]．

ツイートの分類を行う研究として，西田ら [7] は，ある話題についてのツイート群に対して，対象のツイートが圧縮されやすいかどうかで，対象のツイートが，ある話題についてのものであるかを分類する手法を提案している．ツイートの分類に関しては，このような，トピックに着目した手法が主流である [8] [9]．本研究は，投稿されたツイートの話題については考慮せず，ツイートの種類に着目している点に特徴がある．

ユーザ同士の類似性を測る研究としては，山本ら [10] が，ツイート数が一時的に増加する現象である「バースト」に着目し，バーストのタイミングが重なるユーザは，ユーザ同士の興味のあるトピックが重なるとして，これをもとに興味と生活リズムに類似性があるユーザの推定を行う手法を提案している．しかし，本研究では，推薦候補となるユーザとユーザ自身の類似性以外にも，推薦候補となるユーザと，ユーザのフォロイーの類似性が推薦の精度に寄与すると考えている．

3.　フォロイーのツイートパターン

3.1　ツイートの種類

本研究では，ツイートの種類として，以下の 4 つを定義した．ここで，本研究においては，実際の投稿者がどのような思いでツイートしているかは重要でなく，あくまでそれを見ているフォロワーがどう捉えるかが重要であると考えている．

- 情報発信
- 情報要求
- 会話
- その他

それぞれの種類とその意味について，表 1 を参照しつつ説明する．情報発信は，フォロワーに対して情報を発信しているツイートを指し，表 1 の情報発信の項目にあるようなツイートが該当する．このツイートの割合が多いユーザをフォローするのは，そのユーザのツイートの内容が直接ユーザにとって有益である場合が考えられ，たとえばニュース記事を読むように Twitter を使っていると推定できる．

情報要求は，フォロワーに対して何らかの疑問を提示し，答えを求めるようなツイートが該当する（表 1 の情報要求の項目）．このツイートの割合が多いユーザをフォローするのは，そのユーザに対して返ってくるリプライによって自分の疑問を解消したい場合や，自分自身が解答する立場でありたいなどの場合が考えられる．

会話は，誰かと会話する内容のツイートを指す．具体例として，表 1 の会話の項目にあるようなツイートがあげられる．この割合が多いユーザをフォローしている場合，会話の相手とのやりとりそのものを楽しむなど，チャットに近い使い方をするユーザであると考えられる．

その他は，表 1 のその他の項目に示しているような，上記のいずれの種類にも該当しないツイートである．このツイートが多いユーザをフォローしている場合，ユーザは，雑多なツイートを拾い読みするように，SNS というよりは，気軽に記事が読めるブログのような使い方をしているユーザではないかと考え

表 2　分類基準と優先順位

優先順位	種類	分類基準
高	会話	リプライであり，非公式リツイートでないツイート
↓	情報要求	“?” で終わっているツイート
↓	情報発信	URL が含まれるツイート
低	その他	上の三つのどれにも該当しないツイート

られる．

なお，これらの分類以外にも有用な分類が存在する可能性はあるが，すでにツイートの役割の主要なものは分類できていると考えたため，本手法ではこれら 4 つの分類に従ってツイートの分類を行う．

3.2　ツイート分類の基準

ここで，前節で述べたツイートの種類の分類基準を述べる．まず，情報発信は，URL が含まれるツイートである．これは，特にインターネットにおいて，情報の発信がなされる際，その真偽を確かめることは難しく，情報の発信には同時にその情報源を載せていることが，そのもっともらしさを助けると考えたためである．次に，情報要求であるが，これはツイートの最後の文字がクエスチョンマーク “?” であるものとした．そして，会話はリプライに当たるものが会話であるとした．このとき，広義にはリプライとして扱われる，非公式リツイート(注1) は含んでいない．非公式リツイートは，会話を引用し，あらためて自分の発言としてフォロワー全体に発信するためのものであり，会話としての性格が薄まっていると考えたためである．最後にその他だが，上記の条件のいずれにも含まれないツイートのすべてがこれにあたる．これら 4 つの基準をまとめたものを表 2 に示す．

また，表 1 の会話にあるツイートのように，複数の基準に当てはまるツイートの存在が考えられるが，会話，情報要求，情報発信の順で優先順位が高いものとし，重複は許容していない．これは，たとえば，URL が含まれているリプライの場合，フォロワーに対する情報発信というよりは，リプライの相手に対する会話であると定義した方が妥当だと考えたためである．また，他に有用な分類としてどのようなものがあるかの検討，および，現在の分類も含め，より正確な分類基準などについては今後の課題としたい．

3.3　ツイートパターン抽出

ツイートの種類の定義にしたがって，フォロイーのツイートを分類する．そして，各種類ごとに割合を算出し，閾値を設定することで，フォロイーをいくつかのパターンに分類する．こうしてできた，いくつかのパターンのことを，ツイートパターンと呼ぶ．たとえば，本論文で提案しているツイートの種類は 4 種類であり，各種類のツイートを，平均値よりも多めにツイートしているか，少なめにツイートしているかで分けるとすると，2 の 4 乗で 16 のツイートパターンができる．この場合，閾値は平均値である．こうすることで，ユーザによってツイートパターンごとのフォロイーの所属数に差異が見られると考えられ，

(注1)：http://twitguide.net/twitter をはじめよう！/リツイート

表 3　各種類についての割合

フォロイー	情報発信	情報要求	会話	その他
A	6.67%	0%	13.3%	80%
B	18.5%	0.64%	34.4%	46.5%
C	26.4%	1.39%	34.7%	37.5%
⋮	⋮	⋮	⋮	⋮
平均	18%	0.95%	25.7%	39.8%

表 4　ツイートパターン

フォロイー	情報発信	情報要求	会話	その他
A	0	0	0	1
B	1	0	1	1
C	1	1	1	0
⋮	⋮	⋮	⋮	⋮

表 5　リランキング結果例

推薦ユーザ	情報発信	情報要求	会話	その他	パターン所属数	類似度
A	0 (7.53%)	0 (0.68%)	1 (61.6%)	0 (30.1%)	10	0.78
B	0 (7.03%)	0 (0%)	1 (82.3%)	0 (10.6%)	10	0.51
C	1 (97.7%)	0 (0%)	0 (0.57%)	0 (1.72%)	8	0.68
⋮	⋮	⋮	⋮	⋮	⋮	⋮

その差異が「Twitter の使い方の違い」を表していると考えた．このことから，所属数の多いツイートパターンと，ユーザ推薦の候補となっているユーザとの類似度により，ユーザ推薦を実現していく．表 3 に，フォロイーごとのツイートを，各種類ごとに割合で出した場合の例と，平均値の例を示す．また，表 4 に，表 3 が与えられた場合のツイートパターンを示す．

ここで，Twitter を利用しているすべてのユーザを利用せずに，フォロイーのみを基準にして分類の閾値を算出し，それを参照しているのは，ユーザが他のすべてのユーザを把握し，フォロイーの選定をしているとは考えにくいためである．

3.4　推薦ユーザの分類とフィルタリング

先述のように分類したフォロイーと同じく，推薦ユーザも分類する．このときの閾値は，推薦ユーザから求めるのではなく，さきほどフォロイーから求め，適用した値をそのまま使用する．これは，ユーザにとっての基準に沿ったユーザ推薦を行うためである．そして，以下の手順に沿って，推薦ユーザのリランキングを行い，結果を出力する．これにより，行われたリランキングの結果は，表 5 のようになる．

（1） ツイートパターンを，所属フォロイーの多い順にランキングしたものから，n（初期値は 1）位のものを取り出す．これをカレントパターンと呼ぶ．カレントパターンに属する推薦ユーザを取り出す．

（2） 同じパターンに属する推薦ユーザに順位をつけるために，カレントパターンに属するフォロイー中の，ツイート種別の平均値をそれぞれ算出する．これらの値と，各推薦ユーザの

ツイート種別の値とのユークリッド距離を算出し，以下の式に従って，類似度を算出する．

$$Similarity = 1 - \frac{euclid}{max(Euclid)}$$

ここで，$euclid$ は「各パターンのツイート種別の平均値と，各推薦ユーザのツイート種別の値とのユークリッド距離」を指し，$max(Euclid)$ は「$euclid$ が取りうる最大値」を指す．

（3） nがツイートパターンの総数に等しいなら終了する．そうでなければ,（4）へ．

（4） nを1増やし，もう一度（1）から繰り返す．

4. 実　　験

この章では，本研究で提案する手法の有効性を確認するための評価実験を行うとともに，実験の手順と結果，考察についても示す．また，先述したように，本手法は，あらかじめ用意された，何らかの興味で統一された推薦ユーザに対し，ランキングを行うものである．よって，今回は，被験者が選んだ，自身が興味を持つハッシュタグ（ツイートに付与できるタグ）を含んだツイートをしているユーザ10人を推薦ユーザとすることで，この要件を疑似的に満たし，実験を行った．

4.1 実験概要

普段からTwitterを利用しているユーザ7人を被験者として，実験を行った．被験者は，自分がフォローしたいと思う順番に，推薦ユーザのランキングを行った．また，ツイートの種類の分布は，ユーザの時系列に従って変化すると仮定し，一定期間のツイートを取得した．具体的には，分類するフォロイーのツイートと，推薦ユーザのツイートを直近1週間分とした．取得するツイートを，量ではなく期間で指定することで，たとえば，ある時間帯にスポーツの中継放送があったとする．その実況を行い，大量にツイートするユーザが多かったとして，普段から実況のみをツイートするユーザの分類と，普段はその他の種類のツイートも投稿しているユーザの分類を分けることができるようになる．実験では，被験者の作成したランキングと，提案手法をもとにしたシステムが出力したランキングがどれだけ近いかを，スピアマンの順位相関係数を用いて評価した．

今回の実験では，先述の4つのツイートの種類に加えて，会話の分類のしかたを3種類用意し，それぞれの場合で，精度に違いが出るかを確かめた．よって，被験者1人につき，実験の結果が3通り存在する．会話の分類は，ユーザに見える会話と見えない会話を区別した場合，ユーザに見える会話のみを使用した場合，見えるものも見えないものも総じて会話とした場合の3つである．そもそも，Twitterにおけるリプライは，発信者と，受信者と，その両方をフォローしているユーザにしか見えないようになっている．そのため，ユーザの，フォロイーがどれだけ会話をしているかという意識に，実際の会話数が伴わない可能性が高い．そこで，先述の3つの会話の区分を用意することによって，ユーザに見えている会話と見えていない会話が，ユーザのフォローの嗜好にどの程度影響するかを確かめる．なお，3種類の実験のすべてで，閾値を平均値とした．これにより，会話の種類を分けた場合では32の，残り2つの場合で16のパターンを使用して実験を行った．

表6　被験者のランキングに対する順位相関係数

被験者	会話を分割	会話を合算	見える会話のみ
A	-0.08	0.28	0.24
B	-0.65	-0.64	-0.33
C	-0.47	-0.57	-0.22
D	-0.55	-0.67	-0.55
E	0.49	0.45	0.27
F	0.53	0.44	0.25
G	0.08	0.08	-0.23

4.2 実験結果と考察

前節で述べたように，被験者の作成したランキングと，本手法によって出力されたランキング間の，スピアマンの順位相関係数を表6に示す．

表6から読み取れることとして，会話を2種類に分割した場合と，会話を合算した場合において，相関の高さが，正の中程度の相関，負の中程度の相関，相関なしに別れていることが分かる．それに対し，見える会話のみを考慮した場合では，相関が全体的に低い．これは，会話に分類されるツイートのうち，見えない会話をすべて切り捨てることで，会話の絶対数が減り，少量の見える会話をしただけで簡単に平均値を超えてしまい，結果的には，正確にパターンとして組み込めるだけの分類でなくなってしまったことによるのではないかと考えられる．

実験後のアンケートにおいて，被験者Bが，新たにフォローする人物はこれまでフォローしてきたユーザとは違ったユーザをフォローしたいと考えた，と回答した．また，被験者Aは，これまでにフォローしたユーザと，違うユーザのフォローは半々くらいを考えている，と回答した．なお，そのほかの被験者の，同じ項目に対する回答は有意でなかった．被験者Bの相関係数は負の相関を示しており，被験者Aの相関係数では相関は見られない．さらに，提案手法では，先述したとおり，システム動作時点での，多数派のフォロイーのパターンをもとにランキングを行っている．このことから，新たにフォローする人物は,（同じ興味についてのユーザの中でも）今までのユーザとは違ったユーザをフォローしたい，と考えているユーザには，現行システムのランキング法を逆に適用したランキングを適用し，どちらのユーザも半々で，と考えているユーザには，現行のランキングの，高い順位のユーザと低い順位のユーザを交互に出力するなどの変更を加えることによって，推薦の精度が上昇すると考えられる．

今回の実験で使用したツイートは，各フォロイーの直近1週間のツイートであった．しかしながら，これによって，1週間のうちにツイートしていなかったフォロイーを無視したことと，1週間のツイート数が極端に少ないユーザに対しては，正しい分類が行われにくいことが，推薦結果に影響していると考えられる．また，1週間のうちに，Twitter社の提供しているTwitterAPIによって取得できる限界数である，3200ツイートに達しているフォロイーが数人しか存在しなかった．このため，

取得するツイートをさらに長く取り，より多くのツイートを分類することで，フォロイー単位の分類精度も上がり，実験の結果が変化することが考えられる．

5. おわりに

本論文では，ユーザのフォロイーをいくつかのパターンに分類することによって，ユーザの Twitter の使い方を推定し，推薦に利用するという，新たなユーザ推薦のモデルを提案した．また，評価実験によって，いくつかの課題や改善点すべき点が明らかになった．今後は，これらの点を考慮し．改善したシステムの開発が必要であると考えている．

具体的な今後の課題としては，ツイートの分類が必要十分であるかの検討があげられる．特に，今回の実験においては，被験者のフォロイーのツイートの種類ごとの割合の中で，情報要求の割合が全体的に低かった．このことから，情報要求そのものが，ユーザの Twitter の使い方に関わらない，必要のない分類である可能性が考えられる．また，既存の分類についても，その分類基準が正確であるとは言い難く，改善が必要であると考えている．たとえば，とある本の著者が,「○月×日に僕の新しい単行本が出ます！」などとツイートしたとする．これは，著者本人が発言しているため，信ぴょう性が限りなく高く，著者をフォローしているユーザにとっては，間違いなく情報発信として捉えられるツイートであるが，現行の基準では URL が含まれていないために，その他のツイートとしてカウントされてしまう．さらに，パターンの閾値についても，今回の実験では単純に平均値としたが，これを変化させることによって出力結果が変化することが考えられ，最も精度が高くなる閾値の模索が必要であると考えている．また，すでに考察で述べたように，ユーザのフォローの意欲が，システムの出力とは真逆であることがあるために，その選択をユーザが行えるようにすることで，より柔軟に推薦を行うことができると考えている．しかしながら，フォロイーが少ないユーザの場合，分類したパターンの中に，ユーザの所属数が 0 のパターンが多く存在することがある．その中から，どのパターンを推薦するべきであるかを判断することは，現在のシステムでは不可能であるため，何か新たな要素によって，これを推測する必要がある．最後に，ユーザの，推薦ユーザに対するフォロー意欲に，ツイートパターンと発信しているトピックのどちらが強く影響するかの評価を行わなければならず，また，複数のトピックに興味を持つユーザについて，トピックによって好むツイートパターンが異なる可能性についての調査も必要であると考えている．

謝　辞

本研究の一部は，平成 26 年度科研費若手研究 (B)(課題番号：24700098) によるものです．ここに記して謝意を表すものとします．

文　献

[1] 竹村光，田島敬史．情報発信の対象範囲に基づく Twitter ユーザの分類．DEIM Forum 2013，B1-6，2013.
[2] 田中淳史，田島敬史．Twitter のフォロー関係のユーザ意図に基づく分類．DEIM Forum 2011，F5-1，2011.
[3] 山下拓也，佐藤晴彦，小山聡，栗原正仁．フォロー関係に基づく Twitter ユーザの分類．情報処理学会 全国大会，2013(1)，pp.107-108，2013.
[4] 康大樹，島田諭，関洋平，佐藤哲司．属性伝搬モデルを用いたマイクロブログのフォロー先推薦法．DEIM Forum 2011，A1-3，2011.
[5] 北村太一，小川祐樹，諏訪博彦，太田敏澄．コミュニケーションに着目した Twitter フォローユーザ推薦．人工知能学会 全国大会，3E1-R-6-5，2012.
[6] 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．電子情報通信学会，IEICE-113，no.150，pp.169-173，2013.
[7] 西田京介，坂野遼平，藤村考，星出高秀．データ圧縮による Twitetr のツイート話題分類．DEIM Forum 2011，A1-6，2011.
[8] 山本修平，佐藤哲司．二段階抽出法を用いた実生活 Tweet のマルチラベル分類．DICOMO2013，pp.64-71，2013.
[9] 塚田文哉，鈴木徹也．Twitter におけるツイート間の反応を考慮した話題分類法．情報処理学会 全国大会，2014(1)，pp.109-110，2014.
[10] 山本修平，若林啓，佐藤哲司．バースト時刻に基づくフォロー先ユーザ推薦手法．WebDB Forum 2014，A-5，2014.