投稿活動の変化に着目した マイクロブログユーザの可視化手法の提案

山口 裕太郎[1]　山本 修平[1]　佐藤 哲司[2]

概要：近年 Twitter に代表されるマイクロブログが注目を集めている．2006 年にサービスを開始した Twitter は，2012 年には 5 億ユーザを突破している．Twitter では，ユーザは，ツイートと呼ばれる長さが 140 文字に制限された記事を投稿できる．本論文では，ユーザがマイクロブログに記事を投稿する時間帯や頻度，リプライやリツイートなどの機能，ツイートの文字数といった，投稿を構成する要素を投稿活動と定義し，ユーザがアカウントを作成した時点から利用を続けていく過程における投稿活動の変化を明らかにすることを試みる．そのために，一定期間におけるユーザの投稿活動を，投稿頻度や投稿文字数などの特徴量でクラスタリングし，クラスタ間の遷移確率を用いて時間縦断分析を行う手法を提案する．提案手法を用いて 1 年間の日本語の Twitter 記事を対象とした分析を行い，分析で得られた結果を踏まえマイクロブログユーザの可視化システムを実装したので報告する．

A Visualization Method of Microblog User Behaviors Based on the Transition in Posting Activities

yutaro Yamaguchi[1]　syuhei Yamamoto[1]　tetsuji Satoh[2]

1. はじめに

近年 Twitter に代表されるマイクロブログが注目を集めている．2006 年にサービスを開始した Twitter は，2012 年には 5 億ユーザを突破している [1]．Twitter では，ユーザは，ツイートと呼ばれる長さが 140 文字に制限された記事を投稿している．投稿に関わる機能として，他のユーザに対する返信（リプライ）や，投稿を引用するリツイート (RT)，自分の投稿に特定の話題を指すタグを付与するハッシュタグが存在する．ユーザはそれらの機能を利用しながら，情報発信や他のユーザとのコミュニケーションを図っている．本論文では，マイクロブログ記事の投稿数，投稿時間帯，リプライや RT などの使用頻度，ツイートの文字数といった，投稿を構成する要素を投稿活動と定義する．ユーザの投稿活動は多様な形態をとると考えられる．例えば，仲間内でのコミュニケーションに Twitter を使用するユーザはリプライを多く使用し，情報発信目的で Twitter を利用するユーザは RT を多く利用したり，長文の記事を多く投稿すると考えられる．

投稿活動はユーザが Twitter の利用を開始した時点から利用を継続する過程で変化すると考えられ，ある 1 つの時点に着目するよりも連続する複数の時点を系列として分析することでユーザの投稿活動の変化を明らかにできると思われる．投稿活動の変化の例として，利用を始めた直後は投稿数やリプライ数が少なかったユーザが利用を続ける内に，知り合いが増えリプライ数が多くなる場合や，反対に，投稿数が多かったユーザでもある時から投稿間隔が長くなり最後は休止にいたる場合などが想像される．また，学生や社会人といったユーザの属性ごとに利用開始から，ある投稿活動にいたるまでに経過する時間が異なる場合が考えられる．

本論文ではユーザの投稿活動の変化を明らかとするために，投稿活動の時間縦断分析手法を提案し，約 1 年間の日本語ツイートを対象に分析を行う．投稿活動の変化を分析

[1] 筑波大学図書館情報メディア研究科
Graduate School of Library, Information and Media Studies University of Tsukuba
[2] 筑波大学図書館情報メディア系
Faculty of Library, Information and Media Science, University of Tsukuba

することで，ユーザがアカウントを作成後に Twitter の利用を開始してから，利用を休止するまでのライフサイクルを解明できると期待される．さらに，分析で得られた結果を踏まえマイクロブログユーザの投稿活動を可視化するシステムを実装する．

本稿の構成を以下に示す．まず 2 節で本論文に関連するユーザの行動に関する関連研究について紹介し，本論文の位置づけを明らかにする．3 節で分析方法について説明する．4 節，5 節で分析結果および考察を述べ，6 節で可視化システムの概要を示す．7 節でまとめと今後の課題について述べる．

2. 関連研究

マイクロブログユーザの行動に関する特性としてツイートの内容，フォロー・フォロワーネットワーク，リプライ，RT の情報などを使用した研究が知られている．Chalmers ら [2] はリプライと，非リプライツイートのそれぞれに対して，投稿間隔と投稿頻度を分析している．分析の結果，リプライツイートと非リプライツイートでは投稿間隔が異なると報告している．Yang ら [3] は情報拡散構造の観点から Twitter とブログとを比較している．1 ヶ月の投稿回数が 30 回以下のユーザは，ブログよりも Twitter の投稿間隔が小さいが，投稿回数が多いユーザほど両者の差は消失していくと報告している．Kwak ら [4] は，Twitter における RT による記事のつながりをツリー構造とみなす RT ツリーを提案し，RT ツリーのシードからの距離とユーザの関係を分析している．島田ら [5] は，Kwak らの RT ツリーを拡張し，非公式な書式を含むリプライおよび RT を用いて，ユーザ間での情報伝播を有向グラフとして分析を行っている．ユーザ全体の 84.4%がリプライや RT をしたことがあり，Twitter を利用する上で他のユーザとの「つながり」を重視するユーザが多いと結論づけている．Masullo ら [6] は，「利用と満足」の観点から Twitter ユーザの活動を分析している．階層的な回帰分析の結果，1 ヶ月または 1 週間あたりの利用時間が長いユーザほど，他のユーザとのつながりによって高い満足を得ているとしている．

このように，投稿に関わる様々な特徴を用いてユーザの行動を分析する研究は数多く知られているが，時間経過に伴って変化する個々のユーザあるいはユーザ群の投稿活動を解明しようとする試みは知られていない．

3. 投稿活動の変化の分析方法

本節では，本論文で提案する投稿活動の変化を分析する手法について述べる．ユーザの投稿活動は投稿数やリプライ数，RT 数などから特徴づけられる．投稿活動には時間的な変化が存在，すなわち，一定期間ごとに分割して抽出したユーザの投稿活動を表す特徴ベクトルが時系列に従い変化すると仮定する．

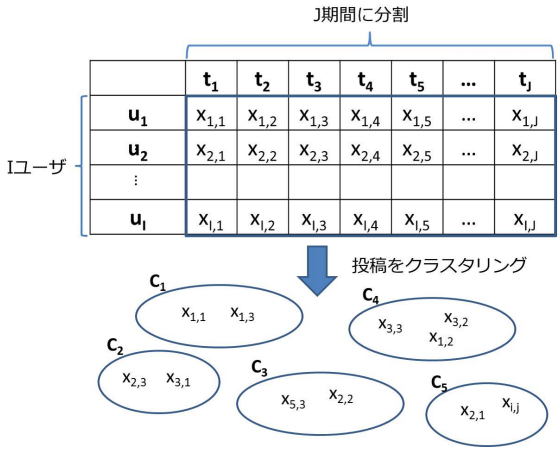


図 1 クラスタリングの概要

本稿においては同一のユーザであっても，異なる期間に抽出された特徴ベクトルは区別して扱う．したがって，特徴ベクトルは分割期間ごとに，分析対象とする全てのユーザから抽出される．特徴ベクトルをクラスタリングすることで，単位時間における投稿活動を分類できる．また，クラスタリング結果を時系列に並べ，ユーザのクラスタ遷移系列とすることで，投稿活動の変化を分析する．なお，Twitter では，同一のユーザが複数のアカウントを使用する場合が生じうるが，本稿では 1 つのアカウントに 1 人のユーザが対応すると仮定し，ユーザとアカウントを区別して扱わない．

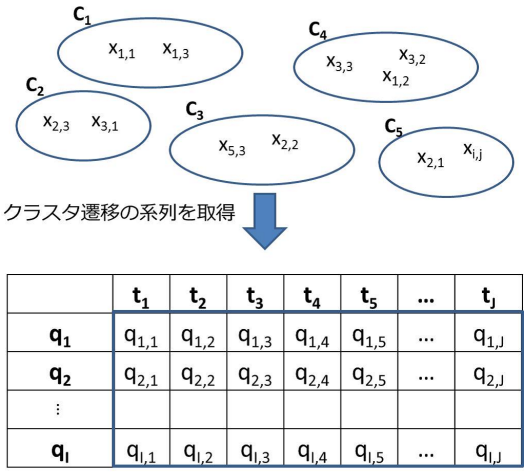
単位時間内のユーザの投稿活動を表す特徴ベクトルの生成法とクラスタリング方法を 3.1 節に，時系列方向の分析方法を 3.2 節に示す．

3.1 投稿活動のクラスタリング

本節では，投稿のクラスタリング方法を説明する．クラスタリングの概要を図 1 に示す．クラスタリングの対象とするのは，一定期間毎に分割したユーザの投稿活動である．あるユーザ u_i が期間 $t_j(1 \leq j \leq J)$ に投稿したマイクロブログ記事集合 $D_{i,j}$ に対し，特徴ベクトル $\boldsymbol{X_{i,j}}$ を作成しクラスタリングを行う．特徴ベクトルは対象ユーザ数 $\boldsymbol{I}$ と分割期間数 $\boldsymbol{J}$ を乗算した数だけ作成される．

特徴ベクトルには，ユーザの投稿活動を特徴づけると考えられる以下の 11 の特徴量を使用する．

(1) ツイートの投稿数 ($\boldsymbol{post}$)
(2) リプライツイートの投稿数 ($\boldsymbol{reply}$)
(3) 異なりリプライユーザ数 ($\boldsymbol{reply\_user}$)
(4) RT 数 ($\boldsymbol{rt}$)
(5) ツイートの平均文字数 ($\boldsymbol{ave\_char}$)
(6) 00:00:00 から 03:59:59 までの投稿数 ($\boldsymbol{post:0-3}$)
(7) 04:00:00 から 07:59:59 までの投稿数 ($\boldsymbol{post:4-7}$)
(8) 08:00:00 から 11:59:59 までの投稿数 ($\boldsymbol{post:8-11}$)
(9) 12:00:00 から 15:59:59 までの投稿数 ($\boldsymbol{post:12-15}$)



	t_1	t_2	t_3	t_4	t_5	...	t_J
q_1	$q_{1,1}$	$q_{1,2}$	$q_{1,3}$	$q_{1,4}$	$q_{1,5}$	...	$q_{1,J}$
q_2	$q_{2,1}$	$q_{2,2}$	$q_{2,3}$	$q_{2,4}$	$q_{2,5}$	...	$q_{2,J}$
⋮							
q_I	$q_{I,1}$	$q_{I,2}$	$q_{I,3}$	$q_{I,4}$	$q_{I,5}$	...	$q_{I,J}$

図 2 クラスタ遷移系列の作成

(10)16:00:00 から 19:59:59 までの投稿数 ($\boldsymbol{post:16-19}$)
(11)20:00:00 から 23:59:59 までの投稿数 ($\boldsymbol{post:20-23}$)

上記の特徴量の全てあるいは一部を使用して，K-means 法によりクラスタリングを行う．K-means 法は特徴ベクトルをクラスタの重心との距離が最小になるように，あらかじめ指定した K 個のクラスタに分割する．K-means 法は，非階層型のクラスタリング手法であり階層型の手法に比べ，計算コストが小さい [*1] ことが特徴である．

3.2 クラスタの遷移系列の作成

クラスタの遷移系列の作成手順を図 2 に示す．3.1 節の手順によって得られたクラスタリング結果を，ユーザごとに時系列に従って並べることで，クラスタの遷移系列を作成する．

ユーザ $\boldsymbol{u_i}$ の投稿活動を表すクラスタの遷移系列 $\boldsymbol{q_i}$ は次のとおり定式化される．ここで $\boldsymbol{q_{i,j}}$ はユーザ $\boldsymbol{u_i}$ の期間 $\boldsymbol{t_j}$ における特徴ベクトルが所属しているクラスタ番号である．

$$\boldsymbol{q_i = (q_{i,1}, q_{i,2}, \ldots, q_{i,J})}$$

次にユーザの投稿活動の遷移を分析する．クラスタリングの結果得られたクラスタを $\boldsymbol{C_l}$ と $\boldsymbol{C_m}$ とすると，$\boldsymbol{C_l}$ から $\boldsymbol{C_m}$ への遷移確率 $\boldsymbol{P_{l,m}}$ を以下の式で算出する．

$$\boldsymbol{P_{l,m} = \frac{n_{l,m}}{\sum_{h=1}^{K} n_{l,h}}}$$

ここで $\boldsymbol{n_{l,m}}$ はすべてのユーザの遷移系列において，クラスタ $\boldsymbol{C_l}$ からクラスタ $\boldsymbol{C_m}$ に遷移した頻度であり，$\boldsymbol{K}$ はクラスタ数である．

得られたクラスタのすべての組み合わせに対して遷移確率を算出し，遷移図を作成することで投稿活動の変化を分析する．

[*1] 後述の第 6 節に示す可視化ツールでリアルタイムな表示を行うためには，高速な手法であることが重要となる．

表 1 使用するデータセット

データセット名	A	B
アカウント作成日	2011 年 11 月 16 日	2011 年 11 月 21 日
分析開始日	2011 年 11 月 16 日	2011 年 11 月 21 日
分析終了日	2012 年 11 月 13 日	2012 年 11 月 11 日
分割数	52	51
対象ユーザ数	8,417	8,837
投稿数	2,802,317	3,394,786
リプライ数	1,186,301	1,484,328

4. 投稿活動の変化の分析

4.1 分析対象

本節では，分析に用いた Twitter 記事の収集方法とデータセットについて説明する．提案手法の評価には 2011 年 11 月から約 1 年間にわたり収集した日本国内で投稿された Twitter の記事 [7] を使用する．ツイートの収集には，Twitter の Search API [*2] を使用した．日本語で記述されたツイートを収集するため，言語に “ja”（日本語）と，日本全域をカバーする位置情報 [*3] とを検索条件として指定した．

データセットの概要を表 1 に，示す．可能な限り長期間にわたるユーザの投稿活動を分析するため，データの収集開始時期にアカウントを作成したユーザを対象とする．また，アカウント作成日による影響を調べるために作成日が異なる 2 つのデータセット A，B を作成した．データセット A は 2011 年 11 月 16 日にアカウントを作成したユーザが投稿した記事であり，データセット B は 2011 年 11 月 21 日にアカウントを作成したユーザが投稿した記事である．なおデータセット A，B ともにユーザの投稿した記事集合の分割期間は 7 日間とした．

データセット A および B における特徴量の最大値を表 2 に示す．投稿数，リプライ数はデータセット A が B よりも大きく，異なりリプライユーザ数とリツイート数についてはデータセット B が大きな値をとった．投稿の平均文字数については同程度の値を示した．各時間帯の投稿数の最大値については，2 つのデータセットともに 4 時台から 7 時台の値が小さく 16 時台から 23 時台の値が大きくなった．

4.2 投稿のクラスタリング結果

本節では，3.1 節で述べた方法を用いて，単位時間におけるユーザの投稿活動をクラスタリングした結果を示す．データセットと特徴量の組み合わせを表 3 に示す．2 種類のデータセットに対して用いる特徴量とクラスタ分割数を変更し，合計 4 種類の組み合わせに対してクラスタリングを行った．

[*2] http://search.twitter.com/search.json
[*3] 兵庫県西脇市を中心とする半径 2,000km 圏内

表 2　特徴量の最大値

データセット名	A	B
post	3,220	2,404
reply	2,363	2,061
reply_user	217	420
rt	435	583
ave_char	177	201
post:0-3	766	825
post:4-7	305	392
post:8-11	616	601
post:12-15	837	503
post:16-19	1,128	736
post:20-23	1,156	871

表 3　特徴量の組み合わせ

組み合わせ名	データセット	特徴量	クラスタ数 (K)
$\boldsymbol{AF_{11}K_{20}}$	A	1〜11（11 次元）	20
$\boldsymbol{AF_{11}K_{10}}$	A	1〜11（11 次元）	10
$\boldsymbol{AF_{5}K_{20}}$	A	1〜5（5 次元）	20
$\boldsymbol{BF_{11}K_{20}}$	B	1〜11（11 次元）	20

クラスタリングには統計解析ツール R[*4] を用いた．また，各特徴量をデータセット中の最大値で除することで値を正規化した．結果には重心の特徴ベクトルのノルムの昇順に，クラスタの番号を付与した．

クラスタに所属する要素数および異なりユーザ数を表 4 から表 7 に示す．表からいずれの組み合わせにおいても，クラスタ 1 に所属する要素数，ユーザ数が最も多いことがわかる．要素数については全要素数の約 73%を占めている．

各組み合わせのクラスタリング結果について，得られたクラスタの重心のレーダーチャートを図 3 から図 6 に示す．図における閉曲線はクラスタの重心の特徴ベクトルであり，多角形の頂点が特徴量の値となっている．いずれの図においても，クラスタ 1 は重心の各次元の値がほぼ 0 となっており，ユーザがツイートを投稿をしていない状態を表している．以降では正規化した値の後の括弧内に正規化前の値を示すこととする．

組み合わせ $\boldsymbol{AF_{11}K_{20}}$ の結果である図 3 において，クラスタ 15，クラスタ 19，クラスタ 20 の重心の投稿数は 0.23(740)，0.56(1803)，0.63(2028) となった．クラスタ 5，クラスタ 6，クラスタ 9，クラスタ 10，クラスタ 13 では重心の投稿数が $\mathbf{0.23 \times 10^{-1}}$(74) から $\mathbf{0.98 \times 10^{-1}}$(315) の範囲の値をとった．それ以外のクラスタにおいては $\mathbf{0.29 \times 10^{-5}}$(0) から $\mathbf{0.56 \times 10^{-2}}$(18) の範囲となった．

組み合わせ $\boldsymbol{AF_{11}K_{10}}$ の結果である図 4 において，クラスタ 4 とクラスタ 9 の重心における投稿数は $\mathbf{0.65 \times 10^{-1}}$(209)，0.30(966) であった．それ以外のクラスタでは投稿数は $\mathbf{0.62 \times 10^{-5}}$(0) から $\mathbf{0.73 \times 10^{-2}}$(23) の間の値であった．

[*4] http://www.r-project.org/

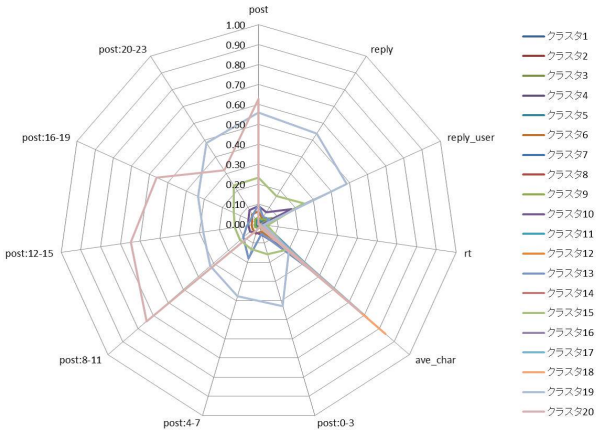


図 3　クラスタの重心：$\boldsymbol{AF_{11}K_{20}}$

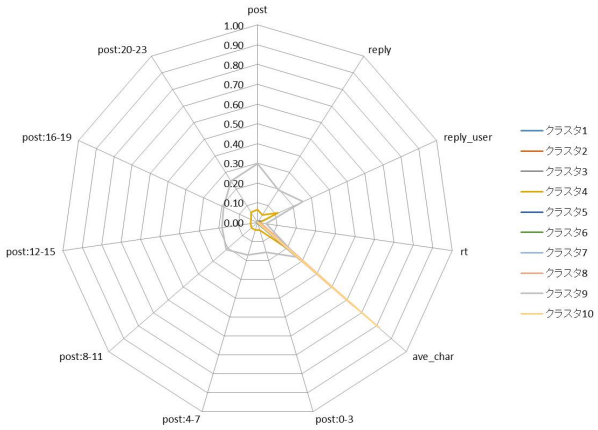


図 4　クラスタの重心：$\boldsymbol{AF_{11}K_{10}}$

組み合わせ $\boldsymbol{AF_5K20}$ の結果である図 5 において，クラスタ 12，クラスタ 18，クラスタ 20 の重心における投稿数の値は 0.22(708)，0.34(1094)，0.63(2028) となった．クラスタ 4，クラスタ 7，クラスタ 8，クラスタ 11，クラスタ 16 では投稿数の値は $\mathbf{0.23 \times 10^{-1}}$(74) から $\mathbf{0.76 \times 10^{-1}}$(244) の間の値となった．それ以外のクラスタは $\mathbf{0.33 \times 10^{-5}}$(0) から $\mathbf{0.55 \times 10^{-2}}$(17) の範囲の値となった．

組み合わせ $\boldsymbol{BF_{11}K_{20}}$ の結果である図 6 において，クラスタ 8，クラスタ 13，クラスタ 17，クラスタ 19，クラスタ 20 の重心の投稿数の値は $\mathbf{0.10}$(240) から $\mathbf{0.65}$(1562) の間をとった．クラスタ 5，クラスタ 10 ではそれぞれ $\mathbf{0.69 \times 10^{-1}}$(165)，$\mathbf{0.40 \times 10^{-1}}$(96) であり，それ以外のクラスタにおいては $\mathbf{0.17 \times 10^{-5}}$(0) から $\mathbf{0.73 \times 10^{-2}}$(17) の間となった．

4.3　クラスタ遷移系列の分析結果

各組み合わせから得られた遷移図を図 7 から図 10 に示す．図において円はクラスタを表し，円の中の数字はクラ

表 4　クラスタの要素数：$AF_{11}K_{20}$

クラスタ	要素数	ユーザ数
1	326,873	8,221
2	7,510	2,216
3	14,513	2,925
4	14,822	3,383
5	5,622	874
6	1,562	324
7	13,804	3,450
8	10,712	3,432
9	2,229	477
10	913	206
11	8,822	3,228
12	7,211	2,701
13	353	53
14	6,686	2,364
15	210	71
16	6,030	2,155
17	4,883	1,917
18	4,888	2,256
19	21	7
20	20	2

表 5　クラスタの要素数：$AF_{11}K_{10}$

クラスタ	要素数	ユーザ数
1	327,703	8,224
2	22,413	3,638
3	27,185	4,055
4	3,763	577
5	1,8282	4,247
6	12,576	3,762
7	10,456	3,021
8	8,440	2,474
9	239	66
10	6,627	2,581

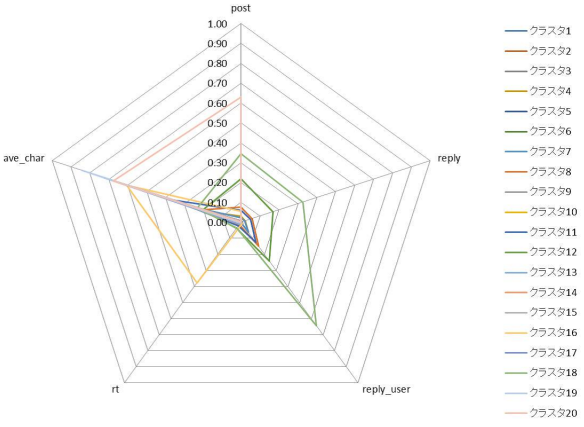


図 5　クラスタの重心：AF_5K_{20}

表 6　クラスタの要素数：AF_5K_{20}

クラスタ	要素数	ユーザ数
1	326,877	8,221
2	7,844	2,237
3	14,542	3,038
4	4,532	738
5	15,280	3,338
6	12,974	3,525
7	4,360	790
8	1,416	289
9	10,981	3,429
10	8,709	3,216
11	683	141
12	232	71
13	7,167	2,661
14	6,593	2,359
15	5,953	2,112
16	115	23
17	4,665	1,912
18	38	16
19	4,703	2,202
20	20	2

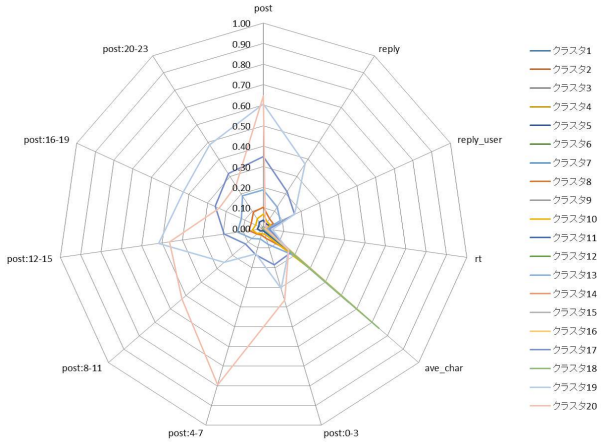


図 6　クラスタの重心：$BF_{11}K_{20}$

スタ番号である．任意の 2 組のクラスタ C_l と C_m 間のパスに付与された数値はクラスタの遷移確率 $P_{l,m}$ である．$P_{l,m}$ はすべてのクラスタの組み合わせについて算出しているが，図においては値が 0.10 以上のパスのみを表示している．

結果から，遷移図は重心のノルムが最も小さいクラスタ 1 に向けて収束する形状を示していることがわかる．いずれの遷移図もクラスタ 1 への遷移確率が 0.10 以上である直鎖と，クラスタ 1 への遷移確率が 0.10 より小さい枝から構成されている．また，クラスタ間にパスが引かれていないことから直鎖を構成するクラスタから枝を構成するクラスタへの遷移確率が 0.10 より小さいことがわかる．遷移図において，クラスタ 1 は他のクラスタに遷移するパスが引

表 7 クラスタの要素数：$BF_{11}K_{20}$

クラスタ	要素数	ユーザ数
1	330,594	8,607
2	6,505	2,084
3	12,984	2,816
4	15,248	3,429
5	5,970	985
6	14,021	3,607
7	12,051	3,674
8	1,767	372
9	9,343	3,445
10	1,406	250
11	8,418	3,237
12	7,746	2,923
13	790	172
14	7,013	2,526
15	6,445	2,316
16	4,936	2,066
17	256	75
18	5,085	2,351
19	83	22
20	77	8

かれない結果となった．クラスタ 1 自身への遷移確率は図 10 においては 0.92 であり，それ以外では 0.93 となった．

組み合わせ $AF_{11}K_{20}$ の結果である図 7 において直鎖は，クラスタ 2，クラスタ 3，クラスタ 4，クラスタ 7，クラスタ 8，クラスタ 11，クラスタ 12，クラスタ 14，クラスタ 16，クラスタ 17 から構成されていた．枝は，クラスタ 5，クラスタ 6，クラスタ 9，クラスタ 10，クラスタ 15，クラスタ 19 から構成されていた．クラスタ 13，クラスタ 18，クラスタ 20 は直鎖と枝には含まれなかった．

組み合わせ $AF_{11}K_{10}$ の結果である図 8 において直鎖は，クラスタ 2，クラスタ 3，クラスタ 5，クラスタ 6，クラスタ 7，クラスタ 8，クラスタ 10 から構成されていた．枝は，クラスタ 4，クラスタ 9 から構成されていた．また，クラスタ 1 以外のすべてのクラスタは直鎖と枝のいずれかに含まれていた．

組み合わせ AF_5K_{20} の結果である図 9 において直鎖は，クラスタ 2，クラスタ 3，クラスタ 5，クラスタ 6，クラスタ 9，クラスタ 10，クラスタ 13，クラスタ 14，クラスタ 15，クラスタ 17 から構成されていた．枝は，クラスタ 4，クラスタ 7，クラスタ 8，クラスタ 11，クラスタ 12，クラスタ 18 から構成されていた．クラスタ 16，クラスタ 19，クラスタ 20 は直鎖と枝には含まれなかった．

組み合わせ $BF_{11}K_{20}$ の結果である図 10 において直鎖はクラスタ 2，クラスタ 3，クラスタ 4，クラスタ 6，クラスタ 7，クラスタ 9，クラスタ 11，クラスタ 12，クラスタ 14，クラスタ 15，クラスタ 16 から構成されていた．枝は，クラスタ 5，クラスタ 8，クラスタ 13，クラスタ 17，クラスタ 19 から構成されていた．クラスタ 10，クラスタ 18，クラスタ 20 は直鎖と枝には含まれなかった．

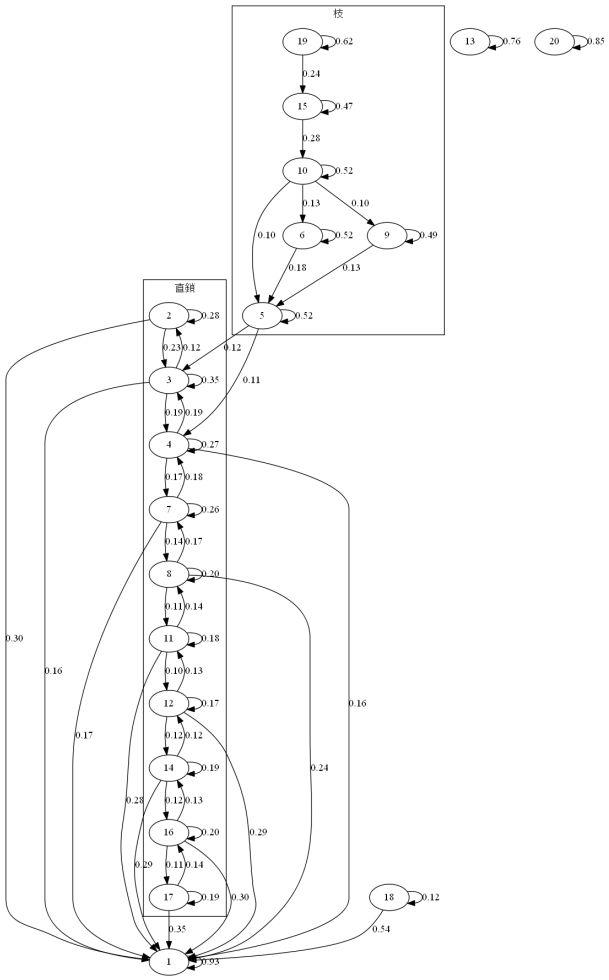


図 7 遷移図：$AF_{11}K_{20}$

5. 考察

5.1 投稿のクラスタリング結果の考察

表 4 から表 7 においてはツイートを投稿していない状態を表すクラスタ 1 に所属する要素数が最も多くなっている．これは，分析対象のユーザの中には Twitter を定常的に利用しているユーザは少なく，多くのユーザが不定期的に投稿を行っているためであると考えられる．

データセット A に対してクラスタの分割数を変化させた，組み合わせ $AF_{11}K_{20}$ と $AF_{11}K_{10}$ の結果に着目すると分割数が 20 の結果である表 4 ではクラスタ 19 やクラスタ 20 のような少数のユーザが所属するクラスタが存在するが，分割数が 10 の結果である，表 5 では存在しない．また，クラスタリング結果の重心の特徴ベクトルを示す図 3 と図 4 を比較すると，分割数が 20 の図 3 ではクラスタ 19 やクラスタ 20 が投稿数が極端に大きいクラスタとして分離できていることがわかる．このことから，クラスタ数 10 よりも 20 の方が適切にクラスタの分割ができることが示唆される．

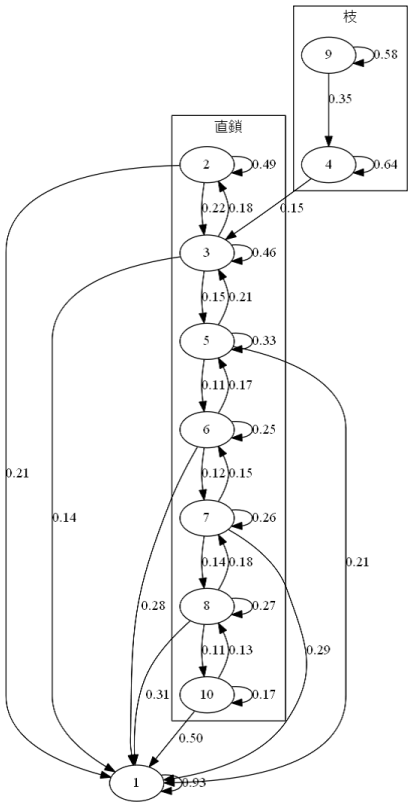


図 8 遷移図：$\boldsymbol{AF_{11}K_{10}}$

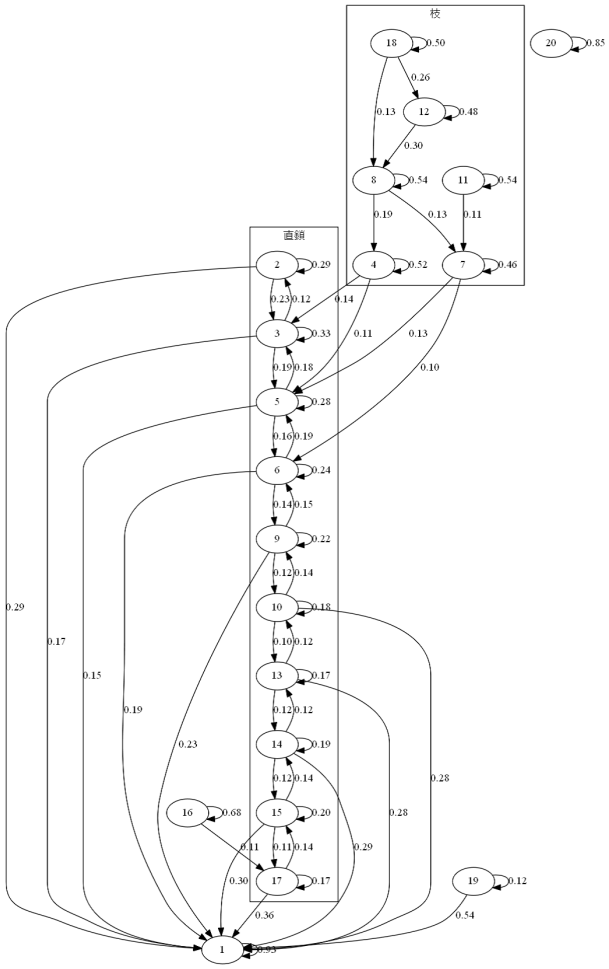


図 9 遷移図：$\boldsymbol{AF_5K_{20}}$

本稿では，投稿活動のクラスタリングにはクラスタ分割数を指定する手法である K-means 法を使用したが，今後はクラスタ分割数を指定しない手法を用いて分析を行う必要がある．

5.2 投稿活動の変化に関する考察

図 7 から図 10 より遷移図の構造が，直接クラスタ 1 に遷移しない枝と直接遷移する直鎖に分かれていることがわかる．枝と直鎖を構成するクラスタの重心を見てみると，4.2 節の結果より枝を構成するクラスタは投稿数が多く，直鎖を構成するクラスタは投稿数が少なくなっている．このことから，Twitter を利用するユーザの投稿活動は投稿数が多い枝の状態を遷移する型と直鎖の状態を遷移する型の大きく 2 種類に分かれると考えられる．枝から直鎖に向かうパスは見られるが，直鎖から枝に向かって遷移するパスが見られないため，投稿数が多い状態から少ない状態に変化することに比べ，投稿数が少ない状態から投稿数が多い状態への変化は起こりにくいことがわかる．

遷移図においてクラスタ 1 は他のクラスタへのパスが引かれずに，自身への遷移確率が大きな値を示す結果となった．これは，Twitter の利用を休止したユーザは再び利用を再開しにくいためであると思われる．

本稿においては，ある状態の直前の状態からの遷移のみを分析したが，今後はそれ以前の状態を考慮した分析を行う予定である．

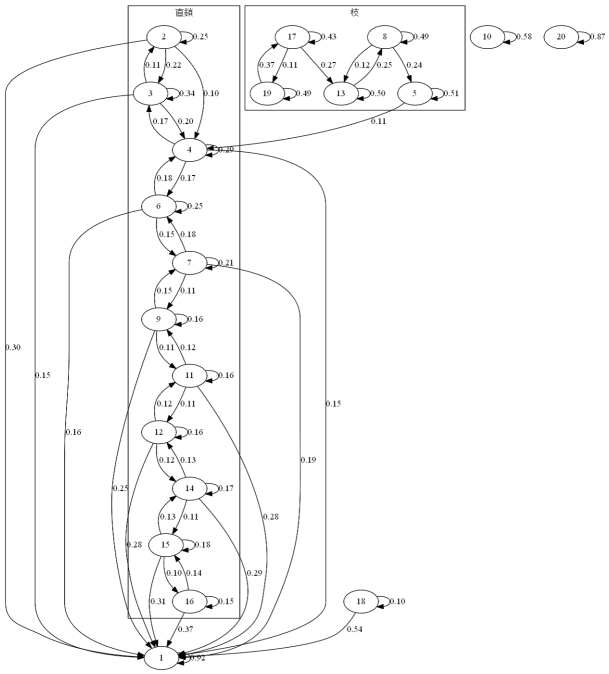


図 10 遷移図：$\boldsymbol{BF11K_{20}}$

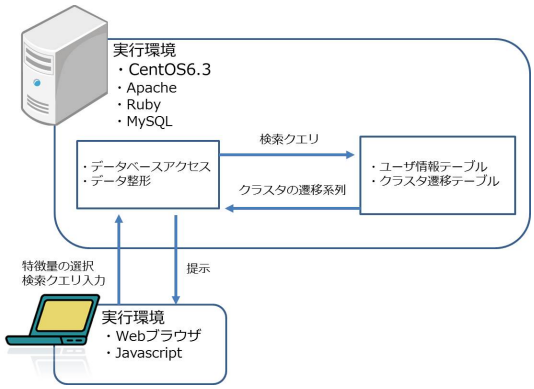


図 11　システム構成

図 12　システムインタフェース

6. 投稿活動の変化の可視化システム

4 節での分析結果を踏まえ，単位時間で分割した個々のユーザの投稿活動がクラスを遷移する様子を分析するために，投稿活動の可視化システムを Web アプリケーションとして実装した．図 11 にシステムの構成を示す．実装環境は，クライアントサイドは javascript を使用し，サーバサイドは Ruby で実装した．データベースは，MySQL を使用した．

図 12 に実行画面を示す．クラスタの遷移系列を色を変えて表示することで Twitter ユーザの投稿活動の変化を可視化する．具体的には，クラスタ番号が小さいクラスタから大きいクラスタをグラデーションを用いて表示する．システムの利用者はクラスタリングに用いる特徴量の選択と Twitter ユーザの自己紹介を記入する bio フィールドに含まれる文字列から分析対象を検索でき，探索的に投稿活動の分析を行うことが可能である．

7. おわりに

本論文ではユーザの投稿活動の変化を明らかとするために，投稿活動の時間縦断分析手法を提案した．1 年間の日本語の Twitter の記事を対象とした分析では，ユーザの投稿活動は，投稿数が多い状態と少ない状態の大きく 2 つに分かれ，投稿数が多い状態からは利用休止状態へ遷移しにくいことが明らかとなった．また，利用休止状態から他の状態への遷移確率は小さく，一度利用を休止したユーザは利用を再開しにくいことが明らかとなった．分析で得られた結果を踏まえマイクロブログユーザの可視化システムを実装した．

今後の課題として，投稿活動の変化と職業や性別などのユーザの属性の関連の分析と，ユーザの投稿活動の変化を表すモデルの精緻化が挙げられる．

謝辞

本研究の一部は，JSPS 科研費 25280110 の助成を受けたものです．

参考文献

[1] TechCruch. Twitter、今年 6 月にユーザー 5 億人超か―ブラジル急成長、ツイート数では日本語が依然英語に次いで 2 位. http://jp.techcrunch.com/archives/20120730analyst-twitter-passed-500m-users-in-june-2012-140m-of-them-in-us-jakarta-biggest-tweeting-city/（参照 2012-10-12）.

[2] Dan Chalmers, Simon Fleming, Ian Wakeman, and Des Watson. Rhythms in twitter. In *SocialCom/PASSAT*, pp. 1409–1414, 2011.

[3] Jiang Yang and Scott Counts. Comparing information diffusion structure in weblogs and microblogs. In William W. Cohen and Samuel Gosling, editors, *Proceedings of the Fourth International Conference on Weblogs and Social Media, ICWSM 2010, Washington, DC, USA, May 23-26, 2010*, pp. 351–354. The AAAI Press, 2010.

[4] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is twitter, a social network or a news media? In *WWW '10: Proceedings of the 19th international conference on World wide web*, pp. 591–600, New York, NY, USA, 2010. ACM.

[5] 島田論, 山口裕太郎, 佐藤哲司. マイクロブログにおける情報伝播距離に着目したユーザプロファイリング. 第 4 回データ工学とマネジメントに関するフォーラム（DEIM Forum 2012）, No. D8-5, 2012.

[6] Gina Masullo Chen. Tweet this: A uses and gratifications perspective on how active twitter use gratifies a need to connect with others. *Computers in Human Behavior*, Vol. 27, No. 2, pp. 755–762, 2011.

[7] 山口裕太郎, 水沼友宏, 山本修平, 島田論, 池内淳, 佐藤哲司. マイクロブログにおける投稿活動に着目したユーザプロファイリング. 第 5 回知識共有コミュニティワークショップ論文集, pp. 1–10, 2012.