マイクロブログにおける投稿活動遷移に着目したユーザのクラスタリング

山口裕太郎[a]　山本修平[b]　佐藤哲司[c]

筑波大学大学院図書館情報メディア研究科

a) *yamaguchi@ce.slis.tsukuba.ac.jp* b) *yamahei@ce.slis.tsukuba.ac.jp* c) *satoh@ce.slis.tsukuba.ac.jp*

概要　近年 Twitter に代表されるマイクロブログの利用が定着してきている．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，投稿活動を構成する要素を抽出し，投稿活動の時間的な遷移に基づいてユーザをクラスタリングする手法を提案する．ここでは，投稿活動を構成する要素のうち，投稿数の変動のパターンに着目し，変動のパターンと他の要素との関係を分析する．具体的には，投稿数の平均値からの差分に基づいた特徴量を使用してユーザをクラスタリングし，クラスタリング結果を解釈するために各クラスタの頻出する遷移パターンを抽出する．1 年間にわたる日本語のツイートを対象に，遷移パターンとユーザの関係について分析を行い，提案法がユーザのクラスタリングに有効であるとの結果を得たので報告する．

キーワード　マイクロブログ，時系列分析，クラスタリング，投稿活動

1　はじめに

近年 Twitter に代表されるマイクロブログの利用が定着してきている．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などが存在する．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．ユーザの投稿活動は，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，ツイートの文字数などで特徴付けられる．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，文字数が 140 文字に近いツイート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，連続する複数の時点を系列として分析することが有効であると思われる．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．

本稿では，投稿活動を特徴づける要素のうち，投稿数の変動に着目する．約 1 年間の日本語ツイートを対象に分析を行い，変動の遷移パターンと投稿数やリプライ，RT などの使用頻度との関係を明らかにすることを試みる．投稿活動の遷移パターンとユーザの関係が明らかになれば，マイクロブログユーザの属性や利用形態の分析のための有効な特徴量の一つとなることが期待される．

本稿の構成を以下に示す．まず 2 節で本論文に関連するマイクロブログユーザの投稿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 節で提案する分析方法について説明する．4 節で分析結果および考察を述べ，5 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投稿に着目した研究は，着目する機能で大別でき，リツイート（RT）・リプライ [2, 3, 4] やツイートの投稿間隔 [5, 6] などが知られている．Kwak ら [2] は，Twitter における RT によるツイート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 [3] 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している．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．Ghosh ら [4] は time-interval と user のエントロピーを用いて RT を分析している．分析の結果，RT は automatic/robotic

activity，newsworthy information dissemination，advertising and promotion，campaigns，parasitic advertisements の 5 つのカテゴリに分類できると報告している．Chalmers ら [5] 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 [6] 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．ユーザの最小の投稿間隔をブログと比較した結果，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．

一方で，Web コミュニティや SNS のユーザのライフサイクルに関する研究も知られている [7, 8, 9]．Danescu-Niculescu-Mizil ら [7] は，Web コミュニティのユーザが使用する言語の変化を 2-gram 言語モデルを用いて分析している．ユーザのライフサイクルはコミュニティの言語に適応する linguisically innovative learning phase と言語の変化を受け入れない conservative phase の 2 段階からなると結論付けている．Dror ら [8] は，質問回答サイトにおいてサービスの利用を停止するユーザを推定している．利用を停止するユーザとそうでないユーザの違いとして，ユーザの質問に対して回答を得られた回数とユーザの回答がベストアンサーに選ばれた回数を挙げている．Kawale ら [9] は，オンラインロールプレイングゲームを対象にユーザ間の社会的影響とゲームへの参加度合いに基づく予測モデルを提案し，利用を停止するユーザの推定を行なっている．

先行研究と比較して，本論文の特徴は，マイクロブログユーザの投稿活動を特徴づける要素のうち，投稿数の変動に着目していること，および約 1 年間に渡る長期間の日本語ツイートを対象に分析していることである．また，ユーザの投稿活動の遷移パターンの分析は，ユーザのライフサイクルの解明にも寄与すると期待される．

3　投稿活動の遷移に着目した分析方法

3.1　分析方法の概要

ユーザの投稿活動の遷移を分析する方法について述べる．分析方法の概要を図 1 に示す．本稿では，ユーザの投稿活動が時系列に従い変化すると仮定する．すなわち，一定期間ごとの投稿から算出した特徴量を時系列に従って並べ，投稿活動の遷移を表す特徴ベクトルを作成する．分析方法は，ユーザのクラスタリングとそれぞれのクラスタに頻出する部分系列の抽出の 2 段階から構成される．

まず，一定期間におけるユーザの投稿数から特徴量を算出し特徴ベクトルを作成する．ユーザの投稿活動を

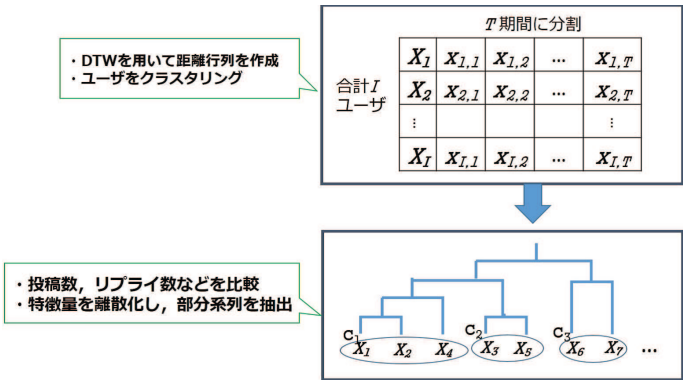

図 1 分析方法の概要

特徴づける要素として，投稿数の変動に着目した特徴量を使用する．クラスタリングを行なうために，ユーザ同士の距離行列を作成する．時間に依存しない類似したパターンを持つユーザを同じクラスタに割り振るという用件を満たす必要がある．そのために，2 つの系列データの距離を時間軸方向に伸縮しながら，最小距離を求める方法である，Dynamic Time Warping(DTW)[10] を距離尺度として用いる．

次に，クラスタリングの結果得られたクラスタを解釈するために，クラスタにおいて頻出する投稿数の変動の遷移パターンを抽出する．その際に実数値で与えられている特徴ベクトルの値を離散化し，整数値の系列データに変換する．部分系列パターンを列挙する手法である PrefixSpan[11] を用いて特徴的な部分系列を取得し，クラスタの意味づけを行なう．

ユーザのクラスタリング方法を 3.2 節で述べ，得られたクラスタからの頻出する部分系列の抽出方法を 3.3 節で述べる．

3.2　ユーザのクラスタリング

一定期間ごとに計算した特徴量をユーザごとに並べることで，時系列に従い変化するユーザの投稿活動を表す特徴ベクトルを作成する．分割期間数を T とした場合のユーザ i の特徴ベクトル X_i を以下に示す．

$$X_i = (x_{i,1}, x_{i,2}, \ldots, x_{i,T})$$

ここで，$x_{i,t}$ はユーザ $i(1 \leq i \leq I)$, の期間 $t(1 \leq t \leq T)$ における特徴量である．投稿数の絶対数が異なるが，上下の変動のパターンが同一の場合が考えられるため，本稿では，投稿数ではなく以下の式で正規化した値を用いる．

$$x_{i,t} = \frac{a_{i,t} - \bar{a_i}}{\sigma_i} \quad (1)$$

$a_{i,t}$ はユーザ i の期間 t における投稿数，$\bar{a_i}$ と σ_i はそれぞれユーザ i の期間 1 から期間 T の投稿数の平均値と標準偏差である．特徴量の値 $x_{i,t}$ は，期間 t のユーザ i

の投稿数を全期間の平均値 a_i と標準偏差 σ_i で正規化した値となる．

ユーザ i から作成された特徴ベクトル X_i と，ユーザ j から作成された特徴ベクトル X_j の，DTW 距離 $dtw(X_i, X_j)$ を全ての i，j の組み合わせに対し算出しユーザ間の距離行列を作成する．

距離行列にクラスタリングを実行し，投稿数の平均値との差分の変動のパターンが類似したユーザのクラスタを作成する．得られたクラスタ間で投稿数やリプライ，RT などの使用割合を比較することで，変動の遷移パターンとその他の指標の関係を分析する．

3.3 頻出部分系列の抽出

本節では，クラスタリングの結果得られたクラスタにおいて，それぞれのクラスタを特徴付ける遷移パターンの抽出法を説明する．式 1 で求めた，各ユーザの特徴量は，実数値で与えられており，部分系列抽出アルゴリズムの適用及び，結果の解釈が困難になることが予想される．そのために，特徴量 $x_{i,t}$ を次の式で離散値 $s_{i,t}$ に変換する．

$$s_{i,t} = sgn(x_{i,t}) * floor(|x_{i,t}|)$$

$$sgn(x_{i,t}) = \begin{cases} 1 & (x_{i,t} > 0) \\ 0 & (x_{i,t} = 0) \\ -1 & (x_{i,t} < 0) \end{cases}$$

ここで，$sgn(x_{i,t})$ は正数に対して 1 を，負数に対し -1 を 0 に対して 0 を返す関数であり，$floor(|x_{i,t}|)$ は実数 $x_{i,t}$ の絶対値に対して小数点以下を切り捨てる関数である．ユーザ p およびユーザ q に対して，特徴ベクトル $X_p = (0.7, 0.0, -1.2)$，$X_q = (1.5, 1.0, 0.0)$ が与えられた場合，$S_p = (0, 0, -1)$，$S_q = (1, 1, 0)$ が離散化後の系列として得られる．離散化を行なうことで，特徴量の値を 0 に近い方向の整数値に丸めている．ユーザ p の期間 1 および期間 2 の投稿数は投稿数の平均値から標準偏差以内であるが，期間 3 においては標準偏差の 1 倍以上 2 倍未満であることを表している．

得られたクラスタの特徴を分析するために，各クラスタ内で頻出する部分系列を抽出する．特徴量を離散化して得られた系列 $S_i = (s_{i,1}, s_{i,2}, \ldots, s_{i,T})$ に対して系列マイニングの手法として知られている PrefixSpan[11] を適用する．PrefixSpan は，長さ 1 の頻出する部分系列から開始して，より長い頻出する部分系列を再帰的に探索するため，候補集合となる部分系列を生成する必要がなく，高速な処理を実現している．また，連続した部分系列に加えて，連続していない部分系列を取得可能である，

表 1 分析対象とするユーザ

条件	2011 年 11 月 16 日にアカウントを作成し同日に 1 回以上投稿したユーザ
対象ユーザ数	1,540
分析開始日	2011 年 11 月 16 日
分析終了日	2013 年 01 月 22 日
分割数	62

4 分析

4.1 データセット

本節では，分析に用いたツイート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 年間収集した日本国内で投稿された Twitter のツイート [12] を使用する．ツイートの収集には，Twitter の Search API [1] 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 [2] とを検索条件として指定した．

分析対象とするユーザを表 1 に示す．長期間の分析を行なうために，収集したツイートの中から 2011 年 11 月 16 日にアカウントを作成し，同日にツイートを 1 度以上投稿したユーザを抽出し分析対象ユーザとする．分析対象である 1,540 ユーザが，2011 年 11 月 16 日から 2013 年 1 月 22 日までに投稿したツイートを分析に用いる．特徴ベクトルの作成のためにツイート集合を 7 日ごとに 62 週に分割し，それぞれの週の投稿数から特徴量を算出する．

4.2 分析結果

投稿活動の遷移パターンが類似したユーザの特徴を分析するために，DTW を用いて作成したユーザの距離行列に階層的クラスタリングを実行し，5 つのクラスタに分割した．クラスタリング方法は Ward 法を選択した．クラスタリングの実行には統計解析ツールの R[3] を使用した．各クラスタに分類されたユーザ数を表 2 に示す．それぞれのクラスタのユーザ数は 200 から 500 の間の値であり，大きな偏りは見られなかった．クラスタ 2 に所属するユーザ数は 5 クラスタ中で最大の 467 ユーザで，クラスタ 4 に所属するユーザ数は最小の 205 ユーザとなっていた．

特徴量として用いた，投稿数の平均値からの変動の遷移パターンと，リプライや RT といった Twitter の機能の使用傾向の関係を観察するために，クラスタに所属するユーザの各週のツイート数（post）に対する，リプラ

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

イ（reply），RT（rt），およびURL（url），ハッシュタグ（hash）を含むツイートの割合を算出した．それぞれのクラスタに所属するユーザの平均値を表3に示す．表中の値は，各ユーザの62週間の平均値をクラスタに所属するユーザで平均した値である．クラスタ1は，ツイートに占めるハッシュタグを含むツイート数が他のクラスタに比べ大きな値となった．クラスタ2の投稿数は，クラスタ5に比べると大きいが，クラスタ1とクラスタ3より小さい値を示している．また，クラスタ2では，リプライ，RT，URLの割合が小さい結果となった．クラスタ3は，クラスタ1と投稿数が同程度であるが，リプライとハッシュタグの割合はクラスタ1より小さい値をとった．クラスタ4は，投稿数が最大のクラスタであり，リプライ，RT，URLを含むツイートの割合も総じて他のクラスタよりも大きい結果となった．クラスタ5は，投稿数が最も小さいクラスタであり，各機能の割合も全て小さい値となった．

それぞれのクラスタのユーザの特徴ベクトルを3.3節の方法で離散化して得られた系列S_iに，PrefixSpanを適用して取得した頻出パターンの上位5件を表4に示す．PrefixSpanの実行にはPrefixSpan-rel[4]を使用した．抽出の際のパラメータである系列長は2以上とした．なお，各週の投稿数の値において，ユーザの62週の投稿数の平均値との差の絶対値が標準偏差以内であることを示す0の出現頻度が大きいため，0をストップワードに設定した．クラスタ5に所属するユーザの特徴ベクトルはストップワードに設定した0が多かったため，条件を満たす部分系列は抽出されなかった．

表4に示す部分系列において，//は前後の要素の間に1つ以上他の要素が入っていることを表している．抽出された部分系列を見てみると，クラスタ1では，クラスタ3に比べ，系列長が長い部分系列が抽出されている．クラスタ2は系列長が長い部分系列は抽出されていない．クラスタ4はそれぞれの週の投稿数の値が大きく正と負の方向に揺れていることがわかる．

クラスタを代表するユーザの特徴量の推移を図2に，離散化した特徴量の推移を図3に示す．なお，クラスタ内で他のユーザとのDTW距離の総和が最小となったユーザを代表的なユーザとして選出した．クラスタ1とクラスタ3は，投稿数の値が平均値よりも大きくなる期間が複数見られた．クラスタ2とクラスタ5は，値のスパイクが見られる．クラスタ5は一度スパイクが見られた後，投稿数は平均値に近い値でほとんど変化していないが，一方のクラスタ2では再度小規模なスパイクが発生している．クラスタ4では他のクラスタに比べて正の方向，負の方向ともに平均値からの投稿数の変位が大きいことがわかる．

[4]http://prefixspan-rel.sourceforge.jp/

表2 各クラスタのユーザ数

クラスタ	ユーザ数
1	274
2	467
3	231
4	205
5	363

表3 各クラスタの62週間の平均値

クラスタ	post	reply	rt	url	hash
1	14.507	0.206	0.080	0.118	0.722
2	1.429	0.026	0.027	0.022	0.544
3	12.370	0.132	0.095	0.079	0.296
4	41.312	0.369	0.108	0.126	0.453
5	0.000	—	—	—	—

4.3 考察

表3において，5つのクラスタは，投稿数が大きいクラスタ4，投稿数が中程度のクラスタ1とクラスタ3，投稿数が少ないクラスタ2，クラスタ5に大別できる．また，投稿数以外にはリプライ，ハッシュタグの割合は全てのクラスタ間で異なっていた．一方で，RT，URLの割合はクラスタ1，クラスタ3，クラスタ4の間では，リプライやハッシュタグの割合ほど大きな違いは見られなかった．図4と図5からクラスタの代表ユーザのRTの割合とハッシュタグの割合の変化を確認すると，RTの割合は，クラスタ4以外はどの週も違いが見られなかったのに対し，ハッシュタグの割合はクラスタ1，クラスタ3，クラスタ4，クラスタ5で投稿の変動と対応した変化が見られた．このことから，投稿数の平均値との差分に基づいた，投稿活動の遷移パターンとの関連が見られる機能と見られない機能があることが示唆される．

表3において，クラスタ1とクラスタ3は投稿数やリプライの割合が類似した傾向を示したが，表4の頻出する部分系列を比較すると，2つのクラスタで異なる傾向を示していることがわかる．このことから，投稿数やリプライの割合だけでは分離が困難であるユーザを，投稿活動の遷移パターンによってクラスタリングできることが明らかになった．

図3の各クラスタの代表ユーザの離散化した特徴ベクトルと表4に示された頻出パターンを比較すると，代表ユーザの特徴量の中に，頻出するパターンが出現していることがわかる．このことから，DTWを距離尺度として用いることで，類似したパターンを持つユーザのクラスタが作成できたと考えられる．一方で，図2におけ

表 4 クラスタで頻出した部分系列

クラスタ 1	頻度	クラスタ 2	頻度	クラスタ 3	頻度	クラスタ 4	頻度
2//2//1//1	80	1//1	101	1//1//2	78	1//1//2//2	72
2//1//1//1//1	74	2//1	68	1//2//1	78	1//2//2//1	72
2//3	72	3//1	60	2//1//1	76	-1//-1//-1//-1//-1//-1//1	70
1//1//1//1//1//1	71	2//2	57	3//2	74	-1//-1//-1//1//1//-1	69
3//2//1	64	5//1	56	2//3	71	1//1//1//1//1//1//1	69

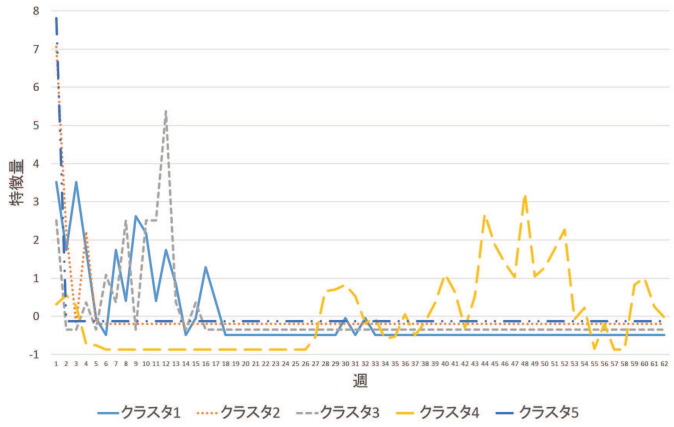

図 2 各クラスタの代表ユーザの特徴ベクトル（実数）

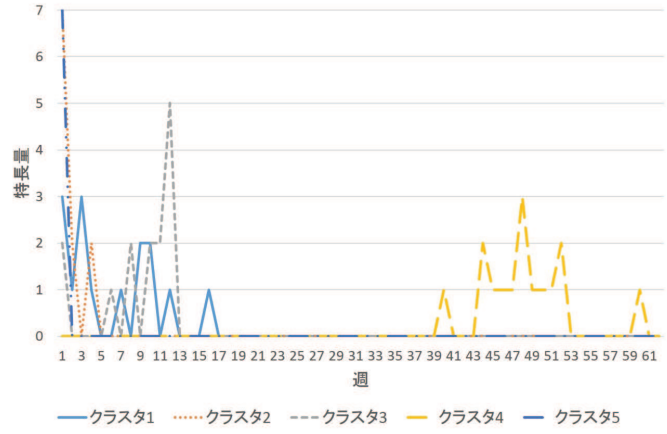

図 3 各クラスタの代表ユーザの特徴ベクトル（離散化）

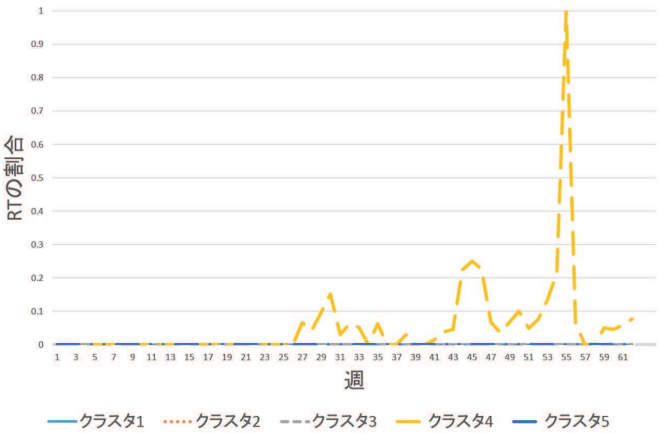

図 4 各クラスタの代表ユーザの RT の割合

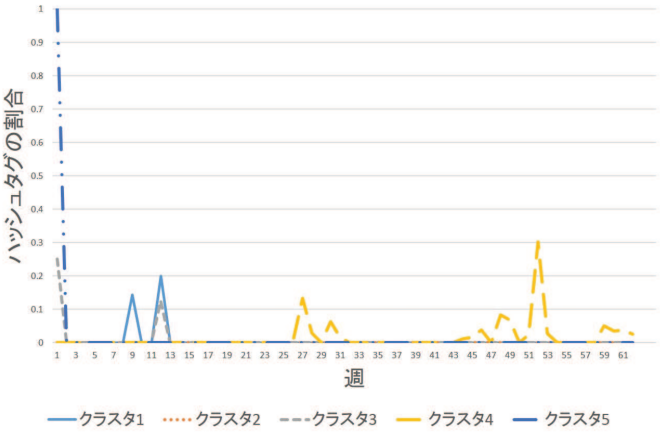

図 5 各クラスタの代表ユーザのハッシュタグの割合

る，クラスタ 4 の 28 週目から 31 週目のスパイクが離散化することで，消失しているため，今後は離散化方法についての検討が必要である．

本稿では，特徴量を作成する際の分割期間を 7 日間に設定したが，時系列の分割の粒度を時間単位といった小さい粒度に設定したパターンを，組み合わせることで詳細な分析が期待できる．

5 おわりに

本稿では，投稿活動を特徴づける要素のうち，投稿数の変動のパターンに着目し，ユーザの投稿パターンと投稿数やリプライ，RT などの使用頻度との関係を明らかにするために，マイクロブログユーザの投稿活動の遷移パターンを用いた分析を行なった．具体的には，各ユーザの平均の投稿数からの差に基づいて計算した，時系列変化を表す特長ベクトルに対して，2 つの系列データの距離を時間軸方向に伸縮しながら最小の距離を求める手法である，DTW を用いて距離を算出し，クラスタリングを行なった．

クラスタリングの結果に対して，離散化した特徴ベクトルに対して，PrefixSpan を適用し頻出部分系列を抽出した結果，DTW 距離を用いることで適切にクラスタリングできたことを確認した．また，得られたクラスタとリプライの割合や RT の割合などを比較した結果，投

稿数の変動の遷移パターンと，リプライやハッシュタグなどのマイクロブログの一部の機能を使用する割合の関連が示唆された．

今後の課題として，時系列の分割の粒度の変更と遷移パターンとユーザの属性やユーザのライフサイクルの関係の分析が挙げられる．

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] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World Wide Web (WWW '10)*, pp. 591–600, 2010.

[3] Satoshi Shimada, Yutaro Yamaguchi, and Tetsuji Satoh. User profiling based on information propagation distance in microblogs (in Japanese). *The 4th Forum on Data Engineering and Information Management (DEIM 2012), D8-5*, 2012.

[4] Rumi Ghosh, Tawan Surachawala, and Kristina Lerman. Entropy-based classification of 'retweeting' activity on twitter. In *Proceedings of KDD workshop on Social Network Analysis (SNA-KDD)*, pp. 143–152, 2011.

[5] Dan Chalmers, Simon Fleming, Ian Wakeman, and Des Watson. Rhythms in twitter. In *Proceedings of the Third IEEE International Conference on Social Computing (SocialCom 2011)*, pp. 1409–1414, 2011.

[6] Jiang Yang and Scott Counts. Comparing information diffusion structure in weblogs and microblogs. In *Proceedings of the Fourth International Conference on Weblogs and Social Media (ICWSM 2010)*, pp. 351–354, 2010.

[7] Cristian Danescu-Niculescu-Mizil, Robert West, Dan Jurafsky, Jure Leskovec, and Christopher Potts. No country for old members: user lifecycle and linguistic change in online communities. In *Proceedings of the 22nd international conference on World Wide Web (WWW '13)*, pp. 307–318, 2013.

[8] Gideon Dror, Dan Pelleg, Oleg Rokhlenko, and Idan Szpektor. Churn prediction in new users of yahoo! answers. In *Proceedings of the 21st international conference companion on World Wide Web (WWW '12)*, pp. 829–834, 2012.

[9] Jaya Kawale, Aditya Pal, and Jaideep Srivastava. Churn prediction in mmorpgs: A social influence based approach. In *Proceedings of the 2009 International Conference on Computational Science and Engineering*, pp. 423–428, 2009.

[10] Hiroaki Sakoe. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. 26, pp. 43–49, 1978.

[11] Jian Pei, Jiawei Han, Behzad Mortazavi-asl, Helen Pinto, Qiming Chen, Umeshwar Dayal, and Mei chun Hsu. Prefixspan: Mining sequential patterns efficiently by prefix-projected pattern growth. In *Proceedings of the 17th International Conference on Data Engineering (ICDE '01)*, pp. 215–224, 2001.

[12] Yutaro Yamaguchi, Yuhiro Mizunuma, Shuhei Yamamoto, Satoshi Shimada, Atsushi Ikeuchi, and Tetsuji Satoh. User profiling based on posting activity in microblogs (in Japanese). *The 5th "Knowledge-share" community workshop*, pp. 1–10, 2012.