

投稿活動遷移に着目した マイクロブログユーザプロファイリングに関する一検討

山口裕太郎[†] 山本 修平[†] 佐藤 哲司^{††}

[†] 筑波大学大学院図書館情報メディア研究科 〒305-8550 茨城県つくば市春日 1-2

^{††} 筑波大学大学院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: [†]{yamaguchi,yamahei,satoh}@ce.slis.tsukuba.ac.jp

あらまし 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短文をリプライやリツイートといった様々な機能を使用して投稿している。本稿では、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といった、投稿活動を構成する特徴量のうち、投稿数の変動に着目する。ユーザの投稿数の時系列変化を用いてユーザをクラスタリングする。ユーザが長期間利用を継続する要因を明らかにするために、利用継続時間の長短で抽出した 2 つのユーザグループに対して、約 1 年間の投稿活動の遷移を分析する。分析結果から、投稿活動の遷移と利用継続時間との関係を議論する。

キーワード マイクロブログ、投稿活動、時系列、クラスタリング

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短文を投稿できる。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）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。このため、ユーザの投稿活動は多様な形態をとると考えられ、マイクロブログ記事の投稿数、投稿時間帯、リプライや RT などの使用頻度などで特徴付けられる。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考えられる。

本稿では、投稿活動を構成する要素のうち、投稿数の変動に着目する。利用継続時間の長短で抽出した 2 つのユーザグループに対して、約 1 年間の投稿活動の遷移を分析し、投稿活動の遷移と利用継続時間との関係を議論する。

本稿の構成を以下に示す。まず 2. 章で本論文に関連するマイクロブログユーザの投稿活動に関する関連研究を紹介し、本論文の位置づけを明らかにする。3. 章で提案する分析方法について説明する。4. 章で分析結果を示し、5. 章で考察する。6. 章で

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関する研究は、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] など着目する機能で大別できる。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。その結果、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、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、両者には投稿間隔に差異があることを明らかにしている。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。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 linguistically 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を推定している．

3. 分析方法

本章では，ユーザの投稿数の変動を特徴量として用いたクラスタリング方法について説明する．まず，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．分析対象とする全期間を等間隔に T 個に区分し，分割したそれぞれの期間 t において直前の期間 $t-1$ との差分を期間 t における特徴量とする．期間 t にユーザ i が投稿したツイート数を $a_{i,t}$ とし，全期間 T 期間にわたる系列 $A_i = (a_{i,1}, a_{i,2}, \dots, a_{i,t}, \dots, a_{i,T})$ が与えられた場合，ユーザ i の特徴ベクトル X_i は以下のようになる．

$$X_i = (x_{i,2}, x_{i,3}, \dots, x_{i,t}, \dots, x_{i,T})$$

$$x_{i,t} = a_{i,t} - a_{i,t-1}$$

ここで， $x_{i,t}$ はユーザ i ($1 \leq i \leq I$) の期間 t ($2 \leq t \leq T$) における特徴量であり，

特徴ベクトル X_i を用いてユーザ同士の距離行列を作成する．ここでは，投稿数の変動をパターンごとに類型化するために Dynamic Time Warping (DTW) [10] を使用する．DTW は，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．そのため，変動の時間軸が異なっている場合でも，パターンが似ているユーザ同士の類似度を評価できる．ユーザ i の特徴ベクトル X_i と，ユーザ j の特徴ベクトル 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距離行列を得る．この距離行列では，投稿の変動パターンが類似しているユーザ相互の距離は小さくなる．

最後に，距離行列に分割最適化クラスタリングの手法である k-medoids 法を適用しクラスタリングを実行する．k-medoids 法の手順は k-means 法と類似しているが，クラスタの代表点に重心ではなく medoid を用いる点が異なる．medoid とはクラスタ内の他の要素との距離の総和が最小となる要素である．k-medoids 法は，k-means 法に比べ，はずれ値の影響が少ないことが知られている．また，k-medoids 法は，要素の距離行列が与えられていれば適用可能である．クラスタ分割数が K 個の場合の，k-medoids 法の手順は次の通りである．

- (1) ランダムに K 個の要素を medoid に設定する．
- (2) 各要素を最も近い medoid に割り当てる．
- (3) クラスタごとに medoid を再選択する．
- (4) medoid の変化が収束または，繰り返し回数が上限を越えれば終了する．それ以外は (2) へ戻る．

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

表 2 グループの概要

グループ名	Long	Short
ユーザ数	500	500
利用継続時間 (日)	最大値	433.98
	平均値	426.07
	最小値	389.59
	標準偏差	10.63
		34.71
		6.19
		0.00
		9.54

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ツイート [11] を使用する．ツイートの収集には，Twitter の Search API (注1)を使用し，言語に”ja”（日本語）と，日本全域をカバーする位置情報(注2)とを検索条件として指定し，日本語で記述されたツイートを収集した．

分析対象とするユーザを表 1 に示す．長期間の分析を行うために，収集したツイートに含まれる 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 件以上投稿したユーザを対象とする．この条件で抽出した 1,540 ユーザを対象に，2011 年 11 月 16 日から 2013 年 1 月 22 日までの投稿活動の遷移を分析する．各ユーザの特徴ベクトル作成にあたっては，ツイート集合を 7 日毎に 62 週に分割し，それぞれの週の投稿数から特徴量を算出した．

上述の 1,540 ユーザの利用継続時間の分布を図 1 に示す．なお，図中の利用継続時間とは，分析期間中の最新の投稿時刻と最古の投稿時刻の差である．図の縦軸は，ユーザ数の累積値を正規化した値を表しており，利用継続時間が 50 日以内と 400 日以上範囲に多くのユーザが存在していることがわかる．

長期間利用を続けるユーザと短期間で利用を休止するユーザの違いを比較するために，利用継続時間が異なる 2 グループを作成した．グループの概要を表 2 に示す．利用継続時間の降順上位 500 ユーザをグループ Long とし，昇順上位 500 ユーザをグループ Short とした．グループ Long は約 1 年以上（最長 433.98 日）利用を継続していたユーザで，グループ Short は 1 ヶ月以内（最長 34.71 日）に利用を取りやめたユーザである．

4.2 クラスタ数の決定

k-medoids 法を適用するにあたり，クラスタ数を決定する必要がある．そこで，クラスタ数を 1 から 30 まで変化させ，それ

(注1): <http://search.twitter.com/search.json>

(注2): 兵庫県西脇市を中心とする半径 2,000km 圏内

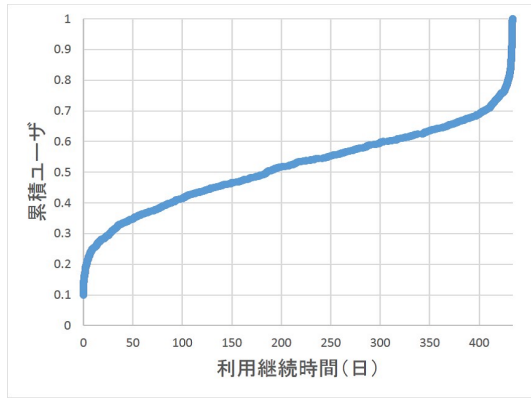


図 1 利用継続時間の分布

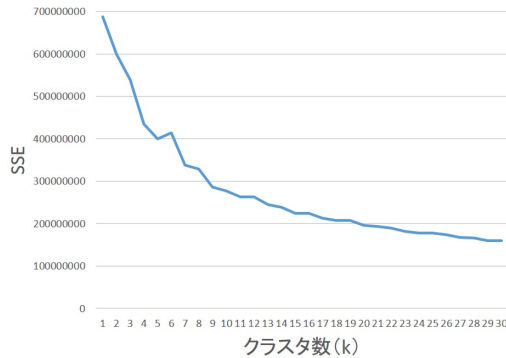


図 2 SSE の推移 (Long)

それぞれのクラスタ数で式 (1) に示す Sum of Squared Error(SSE) を算出した。

$$SSE_K = \sum_{k=1}^K \sum_{\forall X_i \in C_k} dist(X_i, \mu_k)^2 \quad (1)$$

ここで、 K はクラスタ数、 C_k はクラスタ k に含まれる特徴ベクトルの集合であり、 μ_k はクラスタ k の medoid である。また、特徴ベクトル X_i と X_j 間の距離を返す関数を $dist(X_i, X_j)$ とした。SSE の変化が小さい点を最適なクラスタ数とする。

Long および Short の SSE_k の推移を図 2、図 3 に示す。グラフの横軸はクラスタ数であり、縦軸は SSE_k の値である。安定した値を得るために、各クラスタ数において、10 回試行した平均値を SSE_K の値として用いた。また、k-medoids 法の繰り返し回数の上限は 1,000 回とした。

図 2、図 3 の結果からグループ Long ではクラスタ数は 5、グループ Short ではクラスタ数は 3 とした。

4.3 クラスタリング結果

グループ Long

3. 章に述べた方法を用いてグループ Long のユーザの特徴ベクトルを 5 つのクラスタに分割した。クラスタに所属するユーザ数を表 3 に示す。クラスタ番号はクラスタに所属するユーザ数の降順に付与した。クラスタ 1 に所属するユーザ数は最大の 238 ユーザであり、クラスタ 5 に所属するユーザ数は最小の 9 であった。クラスタ 3 とクラスタ 4 に所属するユーザ数は同程

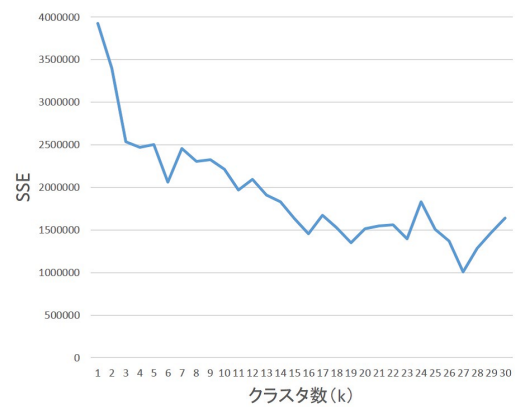


図 3 SSE の推移 (Short)

表 3 各クラスタのユーザ数 (Long)

クラスタ	ユーザ数
1	238
2	162
3	50
4	41
5	9

表 4 各クラスタの 62 週間の平均値 (Long)

クラスタ	post	reply	rt	url	hash
1	6.253	0.138	0.130	0.136	0.042
2	18.238	0.309	0.106	0.105	0.029
3	65.471	0.318	0.144	0.101	0.041
4	63.917	0.409	0.080	0.107	0.031
5	198.817	0.371	0.117	0.068	0.063

度の結果となった。

特徴量として用いた投稿数の変動と、リプライや RT といった Twitter の機能の使用傾向を観察するために、クラスタに所属するユーザの各週のツイート数 (post) に対するリプライ (reply)、RT (rt)、および URL (url)、ハッシュタグ (hash) を含むツイートの比率を算出する。それぞれのクラスタに所属するユーザの平均値を表 4 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 5 は最も投稿数が大きなクラスタである。クラスタ 3 と 4 は投稿数は同程度だが、投稿に占めるリプライの比率はクラスタ 4 の方が大きな値となった。クラスタ 1 は投稿数および、リプライの比率ともに最小のクラスタである。

各クラスタを代表するユーザの特徴ベクトルを図 4 から図 8 に示す。ここで、代表ユーザは、クラスタに属する他のユーザとの距離の総和の昇順 5 ユーザ、すなわち、クラスタの中心付近に位置づけられるユーザである。投稿数が同程度であったクラスタ 3 とクラスタ 4 の特徴ベクトルを比較するとクラスタ 3 に比べて、クラスタ 4 は投稿数の変動が小さい結果となっていた。

グループ Short

3. 章の方法を用いてグループ Short のユーザの特徴ベクトルを 3 つのクラスタに分割した。各クラスタに所属するユーザ

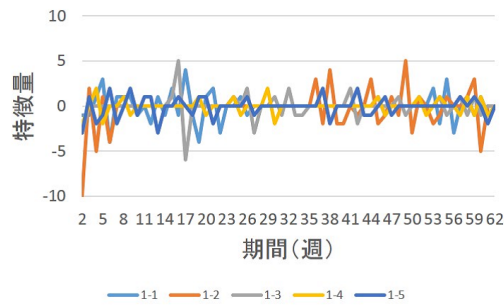


図 4 クラスタ 1 の代表ユーザの特徴ベクトル (Long)

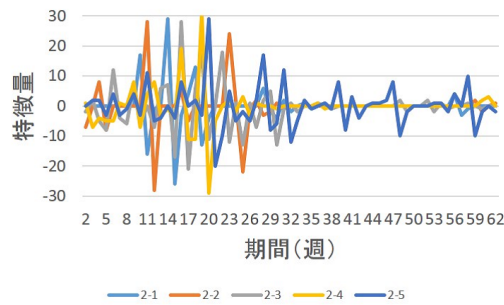


図 5 クラスタ 2 の代表ユーザの特徴ベクトル (Long)

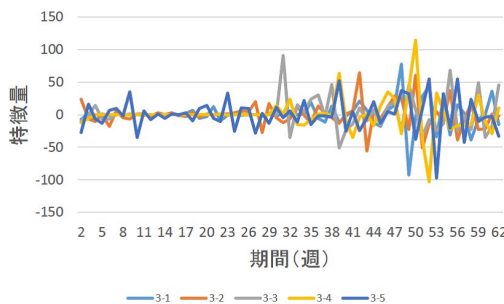


図 6 クラスタ 3 の代表ユーザの特徴ベクトル (Long)

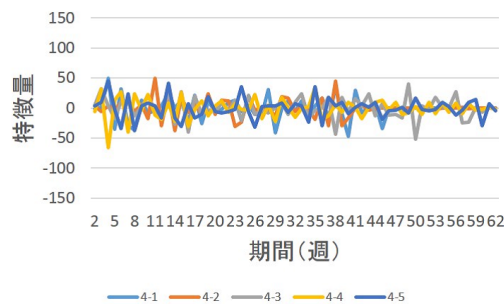


図 7 クラスタ 4 の代表ユーザの特徴ベクトル (Long)

数を表 5 に示す。なお、クラスタ番号は Long と同様に所属するユーザ数の降順に付与した。クラスタ 1 に所属するユーザ数は最大の 412 ユーザであり、クラスタ 3 に所属するユーザ数は最小の 16 であった。Long とは異なり、Short ではユーザの約 80% がクラスタ 1 に所属する結果となった。

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、リプライ (reply)、RT (rt)、および URL (url)、ハッシュタグ (hash) を含むツイートの比率を算出した。それぞれ

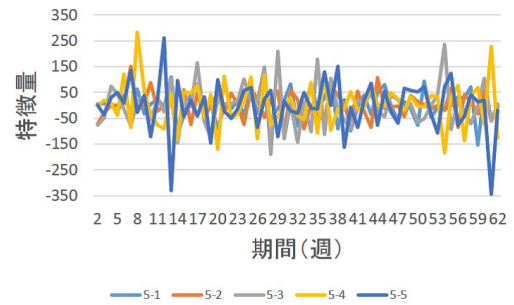


図 8 クラスタ 5 の代表ユーザの特徴ベクトル (Long)

表 5 各クラスタのユーザ数 (Short)

クラスタ	ユーザ数
1	412
2	72
3	16

表 6 各クラスタの 62 週間の平均値 (Short)

クラスタ	post	reply	rt	url	hash
1	0.089	0.004	0.006	0.003	0.001
2	1.899	0.008	0.003	0.003	0.001
3	3.362	0.023	0.003	0.009	0.001

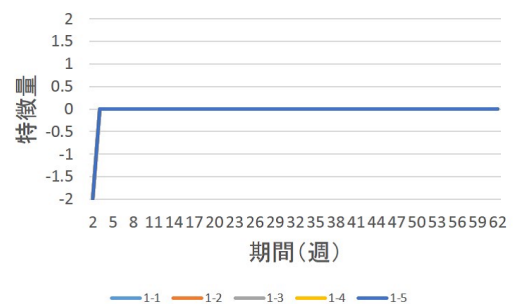


図 9 クラスタ 1 の代表ユーザの特徴ベクトル (Short)

のクラスタに所属するユーザの平均値を表 6 に示す。いずれのクラスタにおいても、グループ Long に比べて小さい値となった。

それぞれのクラスタを代表するユーザを Long の場合と同様の手順で抽出し、その特徴ベクトルを図 9 から図 11 に示す。クラスタ 1 とクラスタ 2 は、変位 (特数量) の大きさは異なるがグラフの形状は類似している。クラスタ 1 とクラスタ 2 は 2 週目に大きく投稿数が変化 (減少) し、それ以降は投稿数の変化が見られない。クラスタ 3 では、利用開始直後に投稿数が変動した後、7 週目以降は安定する結果となった。

5. 考察

グループ Long の分析結果から、図 5 に示すクラスタ 2 は投稿数の変動が大きい期間の後に小さい期間が存在するクラスタ、図 6 に示すクラスタ 3 は変動が小さい期間の後に変動が大きい期間が存在するクラスタ、図 8 に示すクラスタ 5 は一定の変動が全期間を通して存在するクラスタである。そこで、長期間利用を継続する要因を明らかにするために、グループ Long に

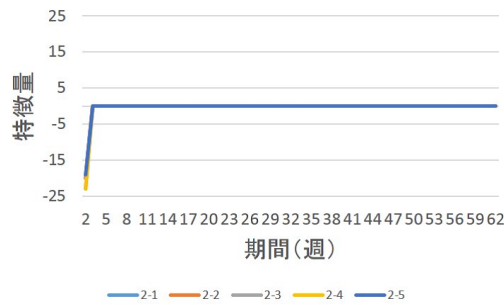


図 10 クラスタ 2 の代表ユーザの特徴ベクトル (Short)

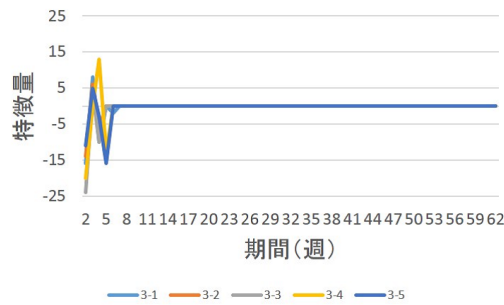


図 11 クラスタ 3 の代表ユーザの特徴ベクトル (Short)

ける投稿数とリプライの比率，RT の比率，URL を含むツイートの比率，ハッシュタグを含むツイートの比率との相関係数を算出する．得られた各クラスタの相関係数を表 7 に示す．相関係数の最大値は 0.2 程度でいずれも強い相関は見られなかった．クラスタ 2，3，5 では投稿数とリプライ数の相関係数が他の機能との相関係数よりも大きな結果となった．

投稿数とリプライの比率の相関係数が他のクラスタに比べ大きかった，グループ Long のクラスタ 2，3，5 において，他のユーザとの距離が最小であったユーザの各週の投稿数とリプライの比率を図 12 から図 14 に示す．各図から投稿数とリプライの比率に相関があることがわかり，クラスタ 2，3，5 の投稿数の変動には，リプライが影響していると考えられる．この結果からリプライを介した，他のユーザとのやり取りが長期間利用を継続する要因と考えられる．

グループ Short の分析結果では，クラスタ 1，2 は 2 週目の変位の大きさは異なるが，3 週目以降の特徴量は 0 となり，類似した形状となっている．これらのクラスタに所属するユーザは，利用開始から 1 週間以内に利用を休止しているユーザだと考えられる．クラスタ 3 では，利用開始直後に変動が見られるが 7 週目以降は投稿数の変動が見られない．このことから，クラスタ 3 は利用開始から約 2ヶ月程度 Twitter を利用した後に，利用を休止するユーザだと考えられる．一定期間利用を継続した後に利用を停止する，クラスタ 3 に所属するユーザは，暫くの間利用を継続した後に，何らかの要因があって利用を休止していることから，このクラスタに属するユーザを詳細に調査することで，休止の要因を明らかにできると期待される．

今回の分析では，長期間利用を継続するユーザの特徴として，投稿数とリプライの比率の相関があることが挙げられた．ここ

表 7 各クラスタの投稿数との相関係数 (Long)

クラスタ	post_reply	post_rt	post_url	post_hash
1	0.090	-0.029	0.046	0.342
2	0.216	-0.039	-0.001	0.023
3	0.234	-0.169	-0.153	-0.076
4	0.050	-0.114	-0.124	-0.090
5	0.215	-0.098	0.015	0.023

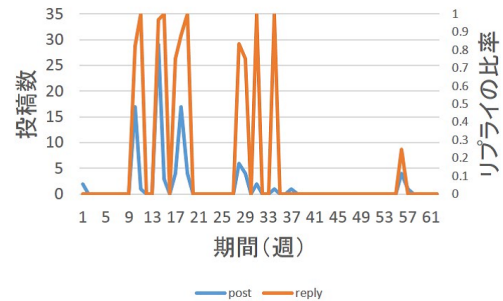


図 12 クラスタ 2 の代表ユーザの投稿数とリプライの比率 (Long)

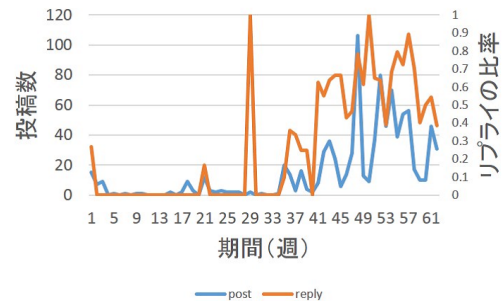


図 13 クラスタ 3 の代表ユーザの投稿数とリプライの比率 (Long)

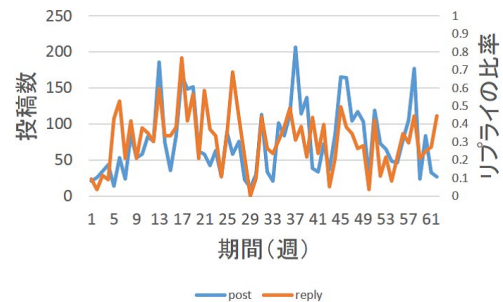


図 14 クラスタ 5 の代表ユーザの投稿数とリプライの比率 (Long)

までの分析では，ユーザが自分から他のユーザに向けてリプライを投稿しているのか，それとも，他のユーザからのリプライに回答する形で投稿しているのか判別ができていないが，リプライは他のユーザをフォローしていることが前提となることから，フォロー関係にある他ユーザの存在が，長期間利用を継続する要因であることは明らかである．今後，ユーザの異なりリプライユーザ数の変化を分析することで，特定のユーザとのみリプライでやり取りをするユーザか，それとも時間の変化とともにやり取りをするユーザ数が増えていくかなどが明らかになると思われ，ユーザのマイクロブログの利用形態をより詳細に明らかにできると期待される．

6. おわりに

本稿では、長期間 Twitter の利用を継続するユーザと、短期で利用を取りやめるユーザの違いを明らかにするために投稿数の変動に着目した分析を行った。

具体的には、利用継続時間が異なる 2 種類のユーザのグループを対象に、ユーザの投稿数の時系列変化を表す特徴ベクトルに対して、DTW 法を用いて距離を算出し、k-medoids 法によりクラスタリングを行った。

分析の結果、利用継続時間が長いユーザの投稿数の変動と投稿に含まれるリプライの比率とに相関があることが明らかとなった。また、投稿活動の時系列変化に着目することで、利用開始直後は投稿数および、リプライの比率も小さいが、利用を継続するうちに両者が増加するユーザの存在も確認できた。

今後の課題として、長期間利用を継続するユーザのリプライの変化の詳細な分析や、ユーザの利用継続時間の推定が挙げられる。

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。

文 献

- [1] TechCrunch. 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? *Proceedings of the 19th international conference on World wide web (WWW '10)*, pp. 591–600, 2010.
- [3] 島田諭, 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) , D8-5, 2012.
- [4] Rumi Ghosh, Tawan Surachawala, and Kristina Lerman. Entropy-based classification of 'retweeting' activity on twitter. *Proceedings of KDD workshop on Social Network Analysis (SNA-KDD 2011)*, pp. 143–152, 2011.
- [5] Dan Chalmers, Simon Fleming, Ian Wakeman, and Des Watson. Rhythms in twitter. *Proceedings of 1st International Workshop on Social Object Networks (SocialObjects 2011)*, pp. 1409–1414, 2011.
- [6] Jiang Yang and Scott Counts. Comparing information diffusion structure in weblogs and microblogs. *Proceedings of the Fourth International Conference on Weblogs and Social Media, (ICWSM-10)*, 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. *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 Szepertor. Churn prediction in new users of yahoo! answers. *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. *Proceedings of the 2009 International Conference on Computational Science and Engineering (ICCSE 2009)*, 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] 山口裕太郎, 水沼友宏, 山本修平, 島田諭, 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.