投稿活動に基づくマイクロブログユーザの
プロファイリングに関する研究

筑波大学
図書館情報メディア研究科

2014年3月

山口裕太郎

目次

第 1 章　序論 1
1.1　背景 1
1.2　本研究の目的とアプローチ 2
1.3　本論文の構成 2

第 2 章　関連研究 3
2.1　マイクロブログユーザの投稿活動に着目した研究 3
2.2　オンラインコミュニティにおけるユーザのライフサイクルに関する研究 . . . 4
2.3　本研究の位置づけ 4

第 3 章　利用開始時期に着目した長期分析 5
3.1　はじめに 5
3.2　利用開始時期に着目した分析手法 5
3.3　分析用データセット 6
3.3.1　ツイートの収集方法 6
3.3.2　収集したツイートの網羅性 7
3.3.3　データセット 11
3.4　長期分析の結果と評価 12
3.4.1　投稿数とリプライ数に着目した分析 12
3.4.2　投稿時刻に着目した分析 12
3.4.3　ツイートの投稿元に着目した分析 18
3.5　考察 18
3.5.1　投稿数とリプライ数に関する考察 18
3.5.2　投稿時刻に関する考察 20
3.5.3　ツイートの投稿元に関する考察 20
3.6　まとめ 21

第 4 章　利用継続時間に着目した短期分析 23
4.1　はじめに 23
4.2　投稿活動の遷移の分析手法 23
4.2.1　クラスタ遷移系列の作成 24
4.2.2　状態遷移図の作成 25
4.3　短期分析の結果と評価 26
4.3.1　データセット 26
4.3.2　投稿活動のクラスタリング結果 27
4.3.3　状態遷移図 29
4.4　考察 29
4.5　まとめ 30

第 5 章　考察 35

第 6 章　結論 36
6.1　まとめ 36
6.2　今後の課題 36

謝辞 37

参考文献 38

発表論文 40

図目次

3.1 分析対象の選定方法 7
3.2 評価実験詳細 9
3.3 時間帯別の投稿数の推移 14
3.4 各時間が一日の投稿に占める割合 15
3.5 時間帯別のリプライ数の推移 16
3.6 各時間が一日のリプライ数に占める割合 17
3.7 各時刻の投稿数に占めるリプライの割合 22

4.1 クラスタ遷移系列の作成方法 24
4.2 状態遷移図の作成例 26
4.3 2 グループの要素数の比率が大きいクラスタの重心の特徴ベクトル 31
4.4 全ユーザの状態遷移図 32
4.5 グループ Long の状態遷移図 33
4.6 グループ Short の状態遷移図 34

表目次

3.1 評価ツイートの網羅性（全体） 8
3.2 収集ツイートの網羅性（投稿数別） 10
3.3 データセット概要 11
3.4 $U_T[t_p, t_q]$ の抽出条件 11
3.5 各グループの投稿数とリプライ数 12
3.6 各グループの投稿元上位 10 件（2006-2007 から 2008） 19
3.7 各グループの投稿元上位 10 件（2009 から 2010） 19
3.8 各グループの投稿元上位 10 件（2011 から new comers） 20

4.1 データセット概要 27
4.2 グループ概要 27
4.3 クラスタの要素数 28

第 1 章

序論

1.1 背景

近年，Twitter[*1] に代表されるマイクロブログが広く普及している．2006 年にサービスを開始した Twitter の，日本国内のユーザ数は 2012 年には 2990 万を突破している [9]．Twitter では，ユーザはツイートと呼ばれる 140 文字以内のメッセージを投稿でき，フォローすることで他のユーザのツイートを閲覧できる．Twitter は，記事を投稿・閲覧する従来のブログ的な側面に加えて，フォローによってユーザ同士が弱くつながる SNS 的な側面を持っている．

ツイートの投稿に関する機能には，他のユーザと会話をする機能であるリプライや，別のツイートを引用するリツイート（RT），ツイートに特定の話題を示すタグを埋め込むハッシュタグなどが存在する．ユーザは，これらの機能を使用してコミュニケーションや情報発信を行っている．

本論文では，投稿に用いる機能や投稿時刻などから特徴づけられるユーザの投稿活動に着目する．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．

ユーザの利用目的を明らかにするためのアンケート調査 [15, 11] も行われている．2010 年に実施された，Twitter の利用目的に関するアンケート調査 [15] では，若年層はリアルタイムのコミュニケーションツールとして利用しており，40・50 代のユーザは情報収集ツールとしての利用が多いと報告されている．2011 年の調査 [11] では，リアルタイムでのコミュニケーションと，趣味の情報を得るために利用する割合が大きいと結論づけている．これらの先行調査報告から，投稿活動とユーザのプロファイルとの間には，いくつかの関連が存在すると考えられ，これらの関連を明らかにすることは，Twitter という新しいメディアの特徴を知る上で

[*1] https://twitter.com/

も重要であるといえる．

1.2　本研究の目的とアプローチ

本研究の目的はプロファイルの異なるマイクロブログユーザ群，あるいは個々のユーザの特徴を投稿活動を通して明らかにすることである．本研究ではユーザのプロファイルとして，「利用開始時期」と「利用継続時間」に着目する．まず，利用開始時期が異なる複数のユーザ群を対象とした分析を行い，マイクロブログユーザコミュニティを構成するユーザの投稿活動の長期的な変化を明らかとする．次に，利用継続時間と投稿活動の関係を明らかにするために，利用継続時間が異なるユーザを対象に分析を行う．そのために，個々のユーザの投稿活動の変化を状態遷移図を用いて分析する手法を提案する．同一の時期にマイクロブログの利用を開始したユーザを，長期間利用を継続するユーザと短期間で利用を辞めるユーザにグループ化し，提案した手法を 2 つのグループに適用することで，投稿活動の変化を俯瞰する．

1.3　本論文の構成

まず 2 章で関連する先行研究について述べ，本研究の位置づけを明らかにする．3 章では，ユーザの利用開始時期に着目した長期分析を行う手法を提案し，収集した実データを用いた評価を行った結果を示す．4 章では，ユーザの利用継続時間に着目した短期分析を行う手法を提案し，個々のユーザの投稿活動の遷移を明らかにする．5 章で，3 章と 4 章で得られた結果について考察する．最後に 6 章で本論文のまとめと今後の課題を述べる．

第 2 章

関連研究

2.1 マイクロブログユーザの投稿活動に着目した研究

マイクロブログユーザの投稿活動に着目した研究は，リツイート（RT）・リプライ [8, 14, 4] やツイートの投稿間隔 [1, 10] など着目する機能で大別することができる．

Kwak ら [8] 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 [14] 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している．

Chalmers ら [1] 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 [10] 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．

2.2　オンラインコミュニティにおけるユーザのライフサイクルに関する研究

Web コミュニティや SNS のユーザを対象にユーザのライフサイクルや行動を分析した研究も知られている [2, 3, 7]．Danescu-Niculescu-Mizil ら [2] 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 [3] 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継続するユーザとでは，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に違いがあることを明らかにしている．Kawale ら [7] 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試みている．

2.3　本研究の位置づけ

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やユーザ群の行動に着目した試みはあまり知られていない．本研究ではプロファイルの異なるマイクロブログユーザ群，あるいは個々のユーザの特徴を投稿活動を通して明らかにすることを目的とする．分析に用いるプロファイルとして，ユーザのアカウント作成時期と利用継続時間を使用し，複数のユーザ群を対象としアカウント作成時期に着目した長期分析と，個々のユーザを対象とし利用継続時間に着目した短期的な分析を行う．

第 3 章

利用開始時期に着目した長期分析

3.1　はじめに

Twitter を利用しているユーザの投稿活動や利用目的は，実社会のイベントや関連サービスなど様々な要因に影響され変化すると考えられる．例えば，Twitter が広く普及する前に利用を開始したユーザには，情報収集目的で利用することが多い，Twitter の普及以降に利用を開始したユーザには，現実の友人とのコミュニケーション目的での利用が多いなどが考えられる．

本章では，Twitter がサービスを開始した当初に利用を開始したユーザから，利用開始直後のユーザまでの投稿活動の違いを分析する．利用開始時期を用いてユーザを層別することで，マイクロブログユーザコミュニティの長期的な変化を解明できると期待される．具体的には，Twitter がサービスを開始した 2006 年から 2012 年までの 6 年間に利用を開始したユーザ集合から，利用開始時期で層別した複数の部分集合を作成し，投稿数やリプライ数，ツイートの投稿元の端末などを比較する．

本章の構成を以下に示す．まず，3.2 節で利用開始時期に着目した分析手法を説明する．3.3 節で分析に用いるデータセットの概要を述べる．3.4 節では分析結果を示し，3.5 節で考察する．3.6 節でまとめを述べる．

3.2　利用開始時期に着目した分析手法

マイクロブログユーザのプロファイルには，年齢や職業，性別などが存在するが，個々のユーザが正確に記述しているとは限らない．ここでは，比較的正確な値が取得でき，かつユーザが利用を開始するきっかけとなったイベントなどの付随的な要因を分析することができる，

Twitter の利用を開始した時期に着目する．

具体的には，アカウントを作成した時期をもとに複数のグループに分類し，それぞれのグループのある期間における投稿活動を比較する．投稿活動を構成する要素として，本論文では投稿数とリプライ数，投稿時刻，ツイートの投稿元を使用する．また，一人のユーザが複数のアカウントを取得する場合があるが本稿では各アカウントをもってユーザと称する．

提案する分析手法は，分析対象とするユーザの選定・グループ化とグループごとの比較の 2 段階からなる．分析対象の選定方法を図 3.1 に示す．ユーザ $u_i(i = A, B, \cdots, E)$ を縦方向に，時間の流れを横方向に示し，各時刻にユーザがツイートを投稿する様子を示している．まず，分析対象期間 T_1 にツイートを投稿したユーザとツイートを収集する．古くから Twitter を利用しているユーザ u_A であっても，期間 T_1 にツイートを投稿していなければ分析対象としない．一方，T_1 にツイートを投稿したユーザの中には，図 3.1 のユーザ u_E のように期間 T_1 の途中から Twitter の利用を始めたユーザも含まれる．このようなユーザは，分析対象期間の全域が利用期間とならないことから除外する．そのための方法として収集したユーザのうちで，期間 T_1 より前の期間 T_0 にもツイートを投稿しているユーザを分析対象のユーザ集合 U とする．この方法では，u_E を除外すると同様に u_B も除外されてしまうが，期間 T_0 を十分に長くすることで問題は軽減される．以上のことから，図 3.1 においては，対象ユーザは u_C と u_D となる．

次に，ユーザ集合 $U = \{u_i|$ 期間 T_0および期間 T_1において 1 回以上の投稿を行ったユーザ $\}$ に含まれるユーザを Twitter の利用開始時期でグループ化する．ここで，利用開始時期すなわち，ユーザがアカウントを作成した時期が t_p から t_q の範囲にあるユーザ集合を $U_T[t_p, t_q]$ とする．Twitter では，ユーザ登録をした順番に各ユーザに id が一意に付与されている．このユーザ id を用いることで，ユーザ集合 U から複数の部分集合 $U_T[t_p, t_q]$ を抽出する．抽出した複数の部分集合間で投稿数とリプライ数，ツイートの投稿時刻，ツイートの投稿元を比較する．以上述べたように本論文で提案する長期分析の手法は，(a) 分析対象期間内のアクティブ（投稿を行っている）ユーザ集合 U を抽出し，(b) ユーザの利用開始時期で U から複数の部分集合 U_T を作成し，(c) 分析対象期間 T_1 における投稿活動を U_T ごとに比較評価する手法である．具体的な評価内容は 3.4 節に示す．

3.3　分析用データセット

3.3.1　ツイートの収集方法

先行研究において Twitter の投稿数が多い言語は英語，日本語，ポルトガル語，インドネシア語，スペイン語だと報告されている [5]．本稿では投稿数が英語についで多く，言語の凝集性が高いことから日本語のツイートを分析対象とする．ツイートは，Twitter の Search

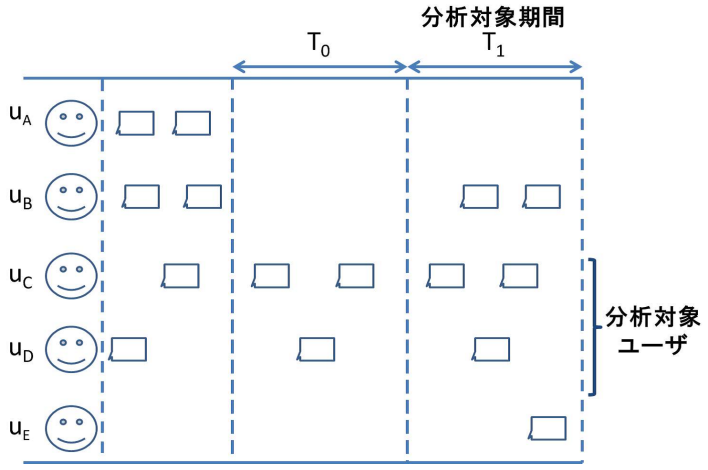


図 3.1　分析対象の選定方法

API *1 を使用して収集した．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．ツイートに付与される位置情報には，ユーザのプロフィール欄に自由記述する「location」情報と，投稿時に GPS 等の値が自動的に付与される「geocode」情報の 2 種類がある．例えば，プロフィールに「茨城県つくば市」と記入しているユーザが，東京スカイツリー（緯度：35.710058 経度：139.810718）でツイートを投稿すると，「location」は「茨城県つくば市」に，「geocode」は「35.710058,139.810718」になる．

位置情報を検索条件とすることで，「geocode」が指定した範囲内にあるツイートが収集できる．「geocode」が付与されていないツイートでは，「location」に記入された情報が参照される．Search API では，「location」が実際の地名等と一致しない場合は，デフォルトで東京とみなしていると思われ，収集の対象となっている．なお，ユーザのフォローに関する情報，お気に入り，およびツイートを非公開に設定しているユーザのツイートは，収集に係る制限が大きいことから，分析の対象には含めていない．

3.3.2　収集したツイートの網羅性

Search API を用いた収集では，１ヶ月に数億件のツイートの収集となることから，収集漏れが発生している可能性が否定できない．そこで，別の収集方法，具体的には，Twitter の REST API *3 を使用した収集方法と比較することで，収集したツイートの網羅性を評価する．REST API では，指定したユーザのツイートを直近 3,200 件まで網羅的に取得できるとされ

*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://api.twitter.com/1/statuses/user_timeline.json

表 3.1　評価ツイートの網羅性（全体）

対象ユーザ数	700
REST API でツイートを収集できたユーザ数	641
REST API で収集されたツイート数	1,028,981
再現率のマクロ平均	0.84
再現率のマイクロ平均	0.86

ている．そこで，ある 1 ヶ月間に 3.3.1 節に述べた方法でツイートを収集できたユーザの中から，投稿数が 100, 200, 500, 1,000, 2,000, 5,000, 10,000 ツイートであるユーザを，各 100 名づつ合計 700 名を抽出し，それぞれのユーザについて，REST API を使ってツイートを収集する実験を行った．投稿数の違う 7 つのグループに分けて，評価ユーザを抽出したのは，投稿数の多寡によって網羅性に違いがあるかを調べるためである．

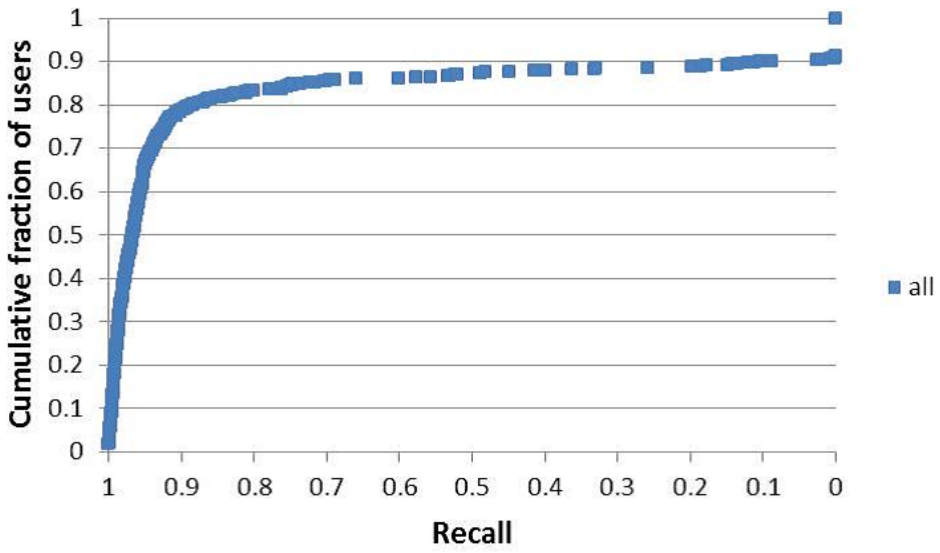
ユーザ単位でのツイート収集は，上記の 700 名の抽出から 3 ヶ月後に実施した．その間に，ツイートを非公開に設定したユーザ，あるいはアカウントが停止されていたユーザが 59 名いた為に，REST API で収集できたユーザは 641 名，総収集ツイート数は 1,028,981 ツイートであった．

網羅性の評価は，情報検索で用いられる一般的な評価尺度である再現率 R を使用する．REST API で収集できる期間がユーザ毎に異なることから，あるユーザ U_i における再現率 R_i は，REST API での収集期間を T_i として，以下の式で求められる．

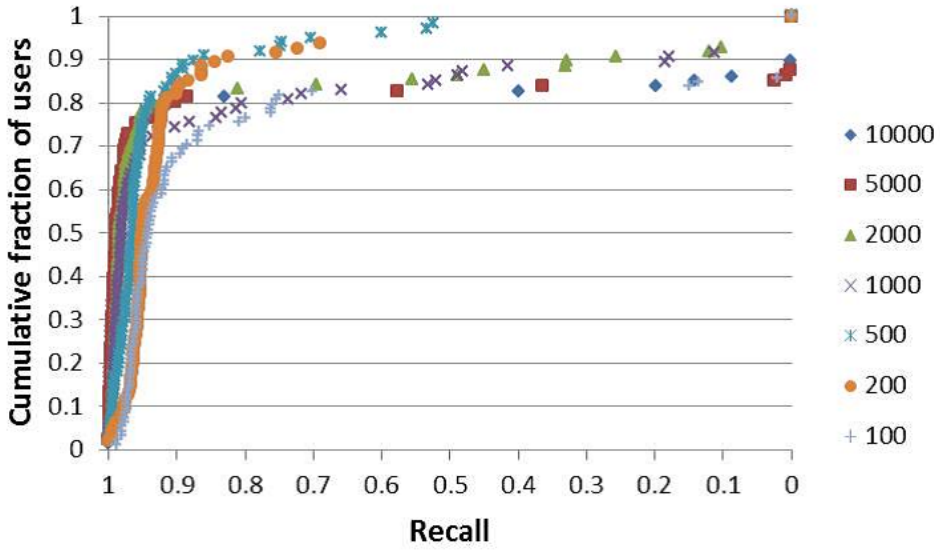
$$R_i = \frac{\text{期間 } T_i\text{に Search API を用いて収集したツイート数}}{\text{期間 } T_i\text{に REST API を用いて収集したツイート数}} \tag{3.1}$$

評価結果を表 3.1，表 3.2 に一覧する．この表で，マクロ平均とは，各ユーザ U_i の再現率 R_i を平均した値であり，マイクロ平均とは，ユーザ毎の区別をせずに，SearchAPI で取得できたツイート数を，REST API で取得できたツイート数で除した値である．各グループに対して，Search API ではツイートを収集できず，再現率が 0 となるユーザが若干名存在したが，当該ユーザも含めて平均値を計算した．表 3.2 の各グループごとの再現率をみると，月間投稿数 200 のユーザのマクロ平均と投稿数 1,000 のマイクロ平均が他のグループに比べ，小さい結果となった．しかし，いずれの再現率も平均して 0.8 を越えており，Search API を使って収集したツイート集合は，各ユーザのツイートを十分な網羅性をもって収集できているといえる．

次に，マクロ平均とマイクロ平均で生じている差異を詳細に分析する為に，投稿数別に再現率の分布を調べた．結果を図 3.2 に示す．図 3.2(a) は全ユーザの再現率の累積グラフ，図 3.2(b) は投稿数別にユーザを分けたときの，それぞれの再現率の変化を表す累積グラフである．



(a) 全ユーザの再現率



(b) 投稿数別ユーザの再現率

図 3.2　評価実験詳細

表 3.2　収集ツイートの網羅性（投稿数別）

1 ヶ月の投稿数	100	200	500	1,000
REST API でツイートを収集できたユーザ数	97	95	96	93
再現率が 0 のユーザ数	14	6	2	8
REST API で収集されたツイート数	49,601	81,160	167,324	163,882
再現率のマクロ平均	0.77	0.88	0.92	0.83
再現率のマイクロ平均	0.76	0.90	0.92	0.82

(a) 投稿数 100 から 1,000

1 ヶ月の投稿数	2,000	5,000	10,000
REST API でツイートを収集できたユーザ数	94	81	85
再現率が 0 のユーザ数	7	10	9
REST API で収集されたツイート数	183,276	159,625	224,113
再現率のマクロ平均	0.85	0.81	0.81
再現率のマイクロ平均	0.88	0.85	0.83

(b) 投稿数 2,000 から 10,000

図 3.2(b) から明らかなように，月間投稿数が 200, 500 のユーザの網羅率は高く，約 90% のユーザは投稿したツイートの 90% を収集できていた．一方，月間投稿数が 100 と 1,000 のユーザは，再現率 0 のユーザ数が他のグループに比べて多く，再現率が低い範囲における累積ユーザ数の上昇が緩やかな傾向を示しているが，その場合であっても，再現率 80% における累積ユーザ数は，ほぼ 80% となっている．

再現率が 0 であったユーザは，REST API でツイートを収集できた 641 名のうち 56 名存在した．それらはユーザ抽出時点では，「location」または「geocode」が収集条件を満たしていたが，評価時点で条件を満たさなくなったユーザである．「geocode」が付与されたツイートは少ないことと，「location」が容易に編集可能であることから鑑みると，収集期間内に「location」を収集条件を満たさないように変更したことが考えられる．実際に，再現率が 0 のユーザを調べたところ，「location」が設定されていない場合や，「location」に数字や顔文字などの日本語と判定できない文字が設定されている場合が確認できた．

再現率が低いユーザは，Search API での抽出時点では，収集条件を満たしていたが，評価に用いた期間 T_i 中に，収集条件を満たしていない時期が存在したユーザであると考えられる．

表 3.3　データセット概要

対象期間	2012 年 5 月 30 日から 6 月 14 日
ユーザ数	2,643,782
投稿数	181,685,501

表 3.4　$U_T[t_p, t_q]$ の抽出条件

グループ名	t_p	t_q
2006-2007	2006 年 8 月 24 日	2007 年 5 月 31 日
2008	2008 年 1 月 1 日	2008 年 6 月 23 日
2009	2009 年 1 月 1 日	2009 年 3 月 29 日
2010	2010 年 1 月 1 日	2010 年 1 月 5 日
2011	2011 年 1 月 1 日	2011 年 1 月 5 日
2012	2012 年 1 月 1 日	2012 年 1 月 4 日
new comers	2012 年 5 月 24 日	2012 年 5 月 29 日

ツイートの公開・非公開の設定や，「location」情報を頻繁に変更するユーザに対しては，網羅的な収集が難しくなっている．

3.3.3　データセット

3.3.1 節に述べた方法で収集したツイートから，3.2 節の方法で分析用データセットを作成した．期間 T_1 は 2012 年 5 月 30 日から 6 月 14 日の約 2 週間とし，T_0 は T_1 の直前となる 2012 年 5 月 14 日から 5 月 29 日とした．表 3.3 に示すとおり，分析対象となるユーザ集合 U の要素数は 2,643,782 ユーザ，投稿数は 181,685,501 ツイートであった．

次にユーザ集合 U からユーザが Twitter の利用を開始した時期にもとづいて複数の部分集合 $U_T[t_p, t_q]$ を作成した．各グループの作成条件を表 3.4 に示す．2006 年から 2012 年の各年の元旦直後に作成された 10,000 アカウントを，部分集合 $U_T[t_p, t_q]$ として抽出した．すなわち，t_p は各年の 1 月 1 日とし要素数が 10,000 ユーザとなるように t_q を設定した．

ただし，2006 年に利用を開始したユーザ数はわずかであったため，Twitter のサービスが開始された時期を t_p とし，それ以降にアカウントを作成した 10,000 ユーザを抽出し，グループ 2006-2007 としている．また，本分析の直前にアカウントを作成した，ユーザ集合 U の中で最も利用開始時期が遅い 10,000 ユーザを new comers として抽出した．

表 3.5　各グループの投稿数とリプライ数

グループ名	投稿数	リプライ数	リプライ率 (%)
2006-2007	985,505	266,463	27.04
2008	967,884	275,610	28.48
2009	919,402	264,541	28.77
2010	747,732	256,389	34.29
2011	609,953	246,173	40.36
2012	618,770	258,704	41.81
new comers	626,571	248,907	39.73

3.4　長期分析の結果と評価

各グループに対して，ツイートの投稿数とリプライ数，一日の投稿数の推移，投稿元ソースの情報について詳細に分析を行った．本節では，各グループの比較結果を示す．

3.4.1　投稿数とリプライ数に着目した分析

各グループごとの投稿数とリプライ数を表 3.5 に示す．投稿数は，古くから Twitter を利用しているグループほど，多い結果が得られた．一方のリプライ数は，どのグループでもほぼ同程度の値であったことから，全投稿数に占めるリプライの割合は，Twitter の利用開始時期が遅いグループほど大きくなる傾向がみられた．

3.4.2　投稿時刻に着目した分析

一日の中での投稿時刻は，ユーザの生活リズムに直結する重要な要素であると考えられる．本章では，一日における投稿数，リプライ数，リプライ率を比較する．

投稿数

一日の中での時間帯別の投稿数の推移を図 3.3（a）に示す．各時間帯 t における平均投稿数 $AveTweets(t)$ は，時刻 t から $t+1$ の直前までの投稿数を評価期間の日数で平均した値である．

$$AveTweets(t) = \frac{1}{N} \sum_{\forall day} Tweets(day, t) \tag{3.2}$$

式（3.2）において，N は分析対象とした日数であり，$Tweets(day, t)$ はある日 day における時刻 t から $t+1$ の直前までの投稿数である．時刻 t を 1 時間単位でとることにすると，$Tweets$(2012-5-30, 0) は，2012 年 5 月 30 日の 0 時 00 分 00 秒から 0 時 59 分 59 秒の間の投稿数である．

平均投稿数は、いずれのグループにおいても，早朝 5 時に最小となり，その後，昼の時間帯で増加傾向を示し，8 時と 12 時頃は局所的なピークとなる．また，13 時以降は投稿数がわずかに低下し，17 時から深夜にかけて再び増加傾向を示す．

図 3.3（b）は，図 3.3（a）の値を各グループにおける一日の投稿数で正規化したグラフである．正規化には次の式を用いて，一日の投稿に占める各時間帯の投稿の割合を求めた．

$$AveTweetsNorm(t) = \frac{1}{N} \sum_{\forall day} \frac{Tweets(day, t)}{Tweets(day)} \tag{3.3}$$

午前 0 時から 7 時までは，各グループともほぼ差異はなくグラフが重なっている．Twitter の利用を開始した時期が早いグループほど午前 8 時から 16 時は値が大きくなった．一方で，利用を開始した時期が遅いグループでは，20 時以降の値が大きい結果となっており，8 時から 16 時とは順序が逆転しているのが特徴的である．

次に式（3.3）の値を平日と休日 [*4] に対して算出した結果を図 3.4 に示す．図 3.4（b) は，休日の投稿であり，休日では，平日に比べて 8 時から 11 時の値が大きかった．また、17 時頃を境に順序が逆転する現象が、平日ほど顕著ではないが休日でも観察された。平日の投稿では 12 時に急激な投稿率の増加がみられるが，休日では，顕著な変動はみられなかった．

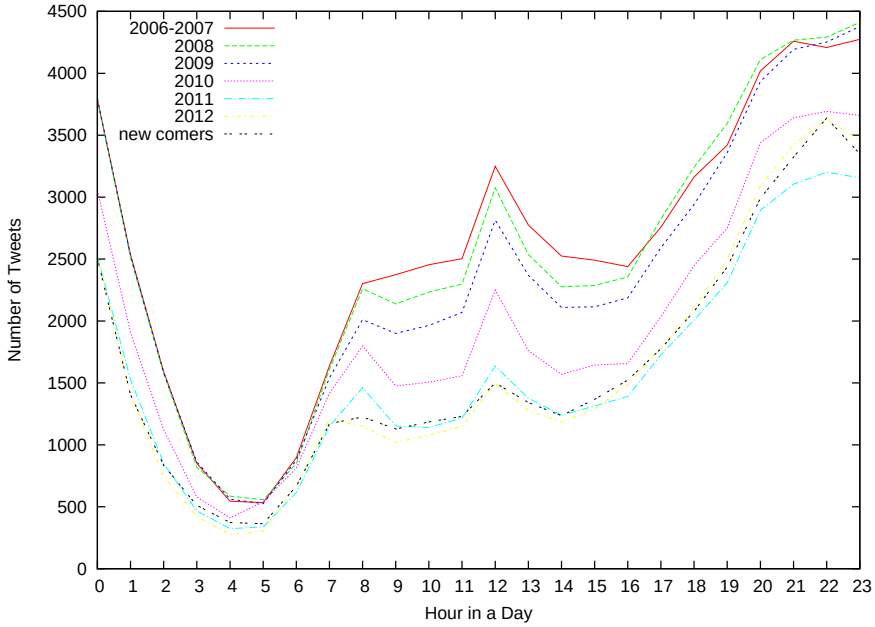
リプライ数

各時刻のリプライ数の平均を図 3.5（a）に，各時刻の値を一日のリプライ数で正規化したグラフを図 3.5（b）に示す．値の算出には，投稿数の場合と同様に式 (3.2) と式 (3.3) を用いた．20 時以降では 2012 と，new comers すなわち，利用開始時期が遅いユーザのグループが他のグループに比べ大きな値を示した．

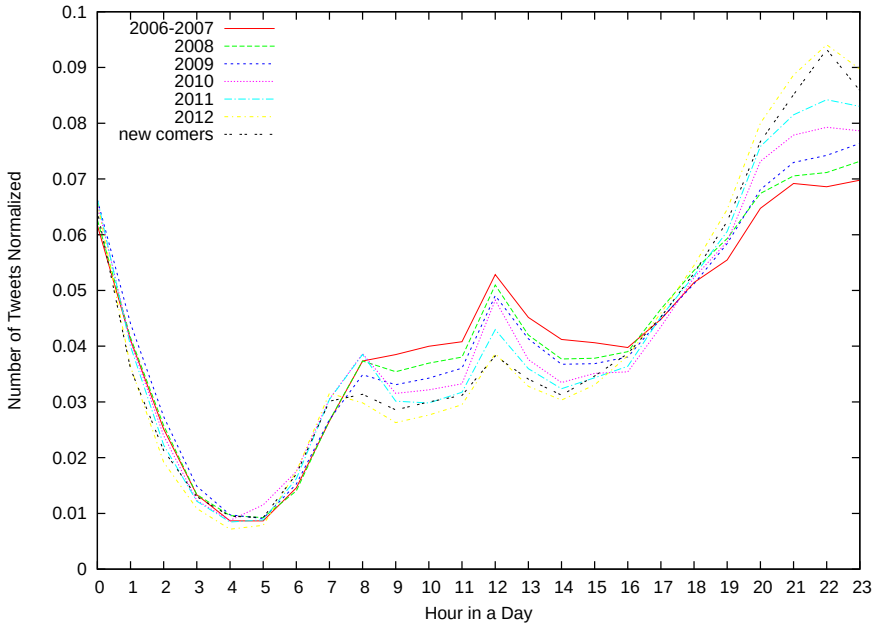
利用開始時期が早いグループでは日中のリプライ数が多いのに対し，利用開始時期が遅いグループでは 20 時以降の値が大きい結果となり，正規化した投稿数の結果（図 3.3（b））と同様に順序の逆転がみられた．

図 3.6 に図 3.5（b）のグラフを平日と休日ごとに描いたグラフを示す．投稿数の場合と同様に，休日では平日に比べ 12 時の値の増加が緩やかな結果となった．

[*4] 国民の休日を含む土曜，日曜を休日とした

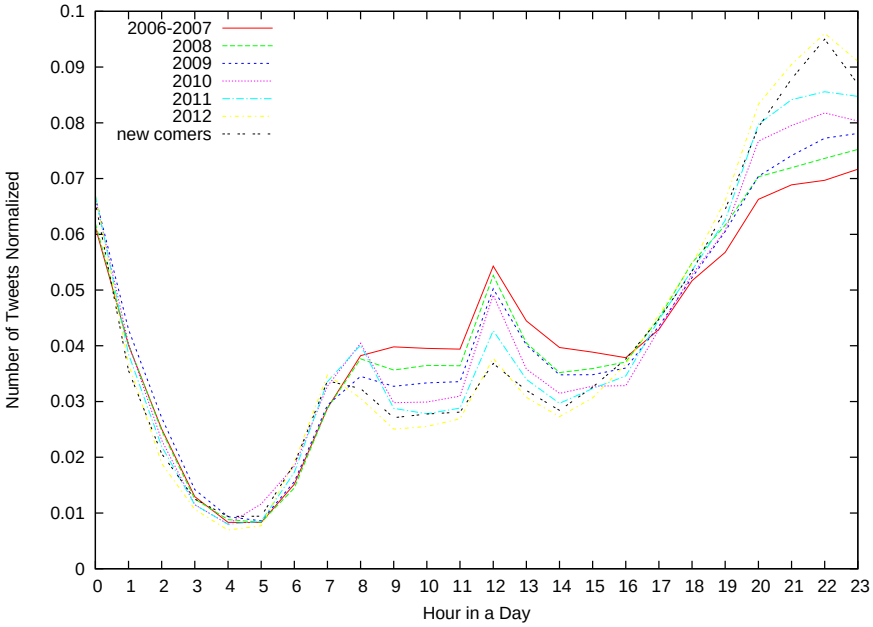


(a) 投稿数

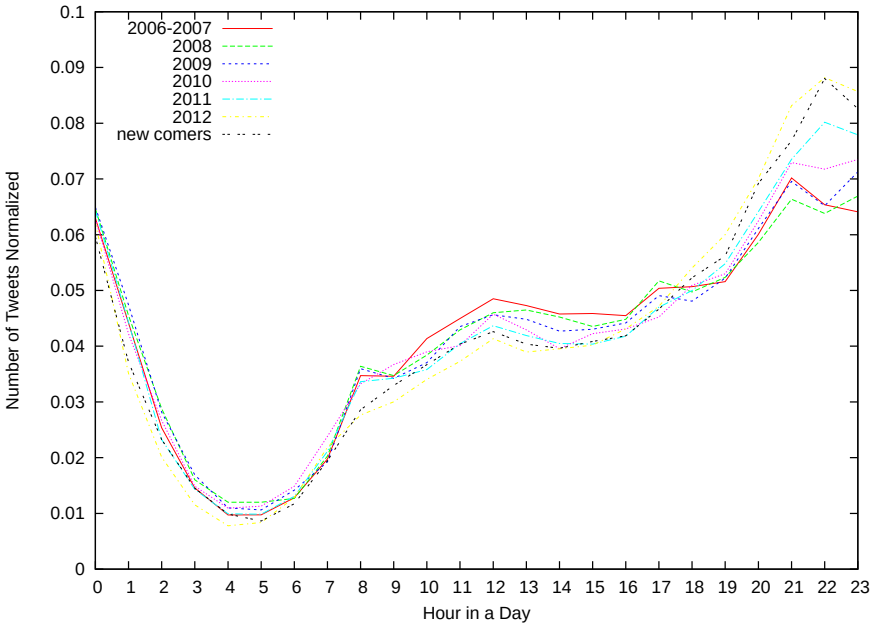


(b) 正規化した投稿数

図 3.3　時間帯別の投稿数の推移

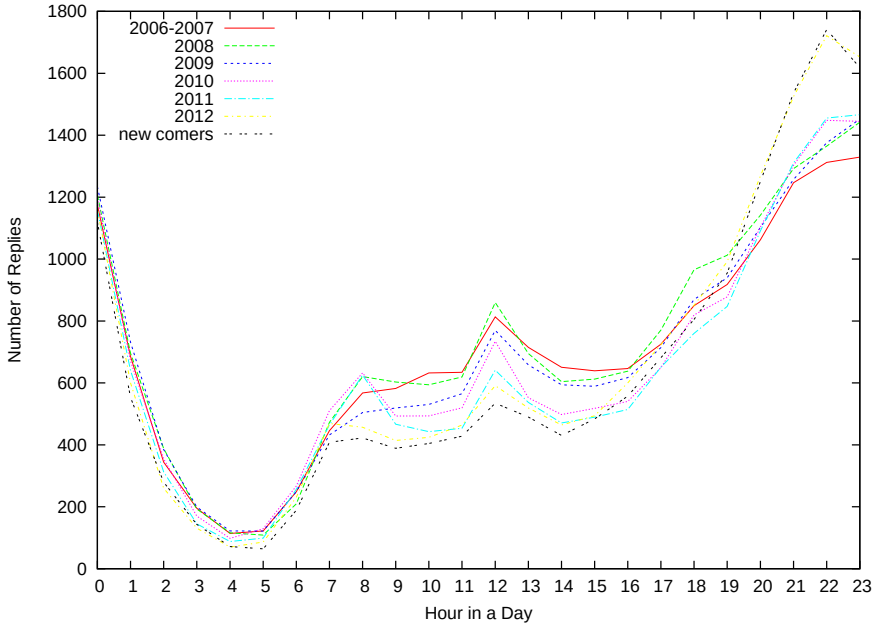


(a) 平日

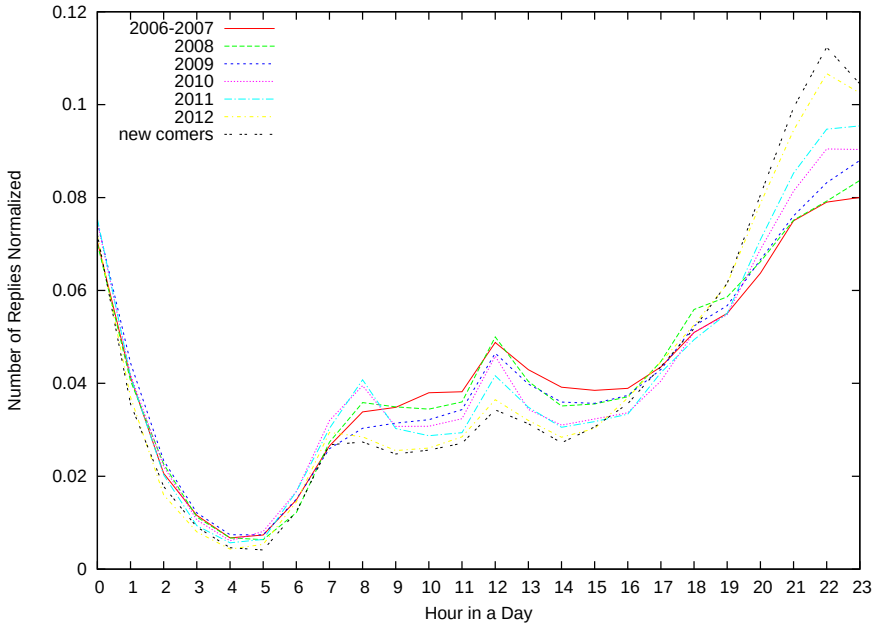


(b) 休日

図 3.4　各時間が一日の投稿に占める割合

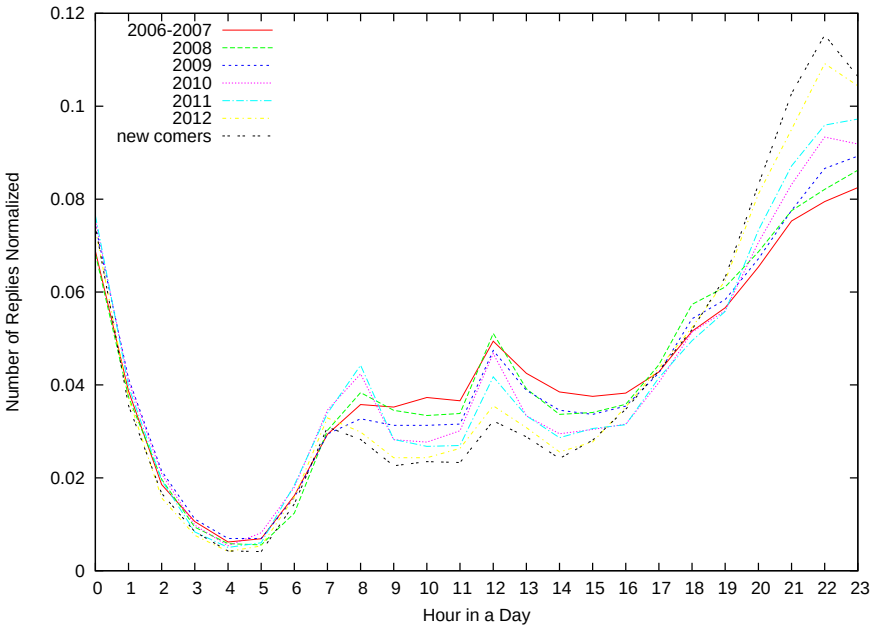


(a) リプライ数

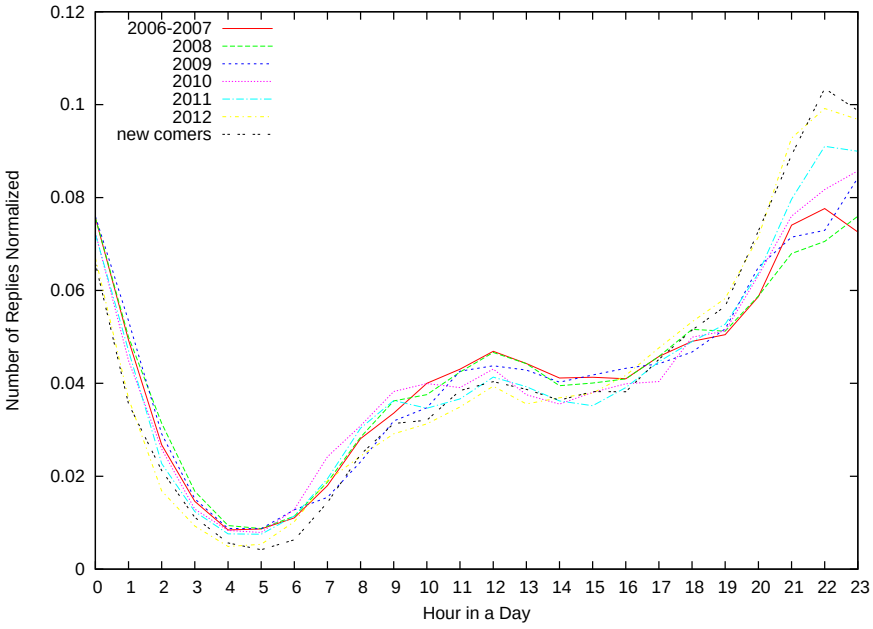


(b) 正規化したリプライ数

図 3.5　時間帯別のリプライ数の推移



(a) 平日



(b) 休日

図 3.6　各時間が一日のリプライ数に占める割合

リプライ率

各時刻の投稿数に占めるリプライの割合を図 3.7 に示す．時刻 t におけるリプライ率 $ReplyRate(t)$ は次式を用いて算出した．

$$ReplyRate(t) = \frac{1}{N} \sum_{\forall day} \frac{Reply(day, t)}{Tweets(day, t)} \tag{3.4}$$

ここで，N は分析対象とした日数，であり，$Reply(day, t)$，$Tweets(day, t)$ はある日 day における時刻 t のリプライ数，およびツイートの投稿数である．

図 3.7 の結果から，どのグループも 20 時以降に高い値を示している．利用開始時期が早い 2006-2007，2008，2009 では，他のグループと比べ，相対的に低い値を示した．投稿数とリプライ数の分析結果において，投稿の占める割合が増加していた 12 時に関しては，リプライ率では増加がみられなかった．その一方で，全てのグループにおいて朝の 6 時に大きな値となった．また，new comers では 6 時から 15 時の間は，リプライ率が 2011 および 2012 よりも低く推移しているが，20 時以降に 2011 と 2012 の値を上回る特徴的な変化を示した．

3.4.3　ツイートの投稿元に着目した分析

ツイートの投稿元の上位 10 件とそのグループ内のツイートに対して，各投稿元からのツイートが占める割合を表 3.6, 表 3.7, 表 3.8 に示す．投稿元の情報は，ツイートに付与されている「source」情報を用いた．2006-2007 から 2010 では，web からの投稿が上位となっている．一方の 2011 から new comers では，Twitter for iPhone や，Twitter for Android といったスマートフォンからの投稿が占める割合が極めて大きくなった．2011 以降では，フィーチャーフォン [*5] 用のクライアントである KeitaiWeb の占める割合が大きいのが特徴的である．

3.5　考察

3.5.1　投稿数とリプライ数に関する考察

3.4.1 節の投稿数とリプライ数に関する分析結果において，利用開始時期が遅いユーザは，リプライの割合が多い結果を示している．2011 年に実施された調査 [11][13] では，Twitter を利用している友人がきっかけで，Twitter の利用を開始したと答えたユーザの割合が多い．このことから，新規ユーザの多くは，友人とのコミュニケーションが目的であり，その結果リプライの割合が多くなったことが推察される．

[*5] スマートフォンではないが通話以外の機能を持つ携帯電話のこと

表 3.6　各グループの投稿元上位 10 件（2006-2007 から 2008）

2006-2007	2008
Echofon(11.59%)	web(11.70%)
web(11.22%)	Echofon(8.38%)
YoruFukurou(7.74%)	twicca(5.99%)
TwitterforiPhone(5.29%)	Tween(5.56%)
twicca(4.02%)	YoruFukurou(5.25%)
TweetbotforiOS(3.73%)	TwitterforiPhone(4.80%)
HootSuite(3.10%)	SOICHA(4.07%)
TweetDeck(3.04%)	Janetter(3.09%)
Tween(3.00%)	TweetbotforiOS(3.05%)
SOICHA(2.88%)	TweetButton(2.90%)

表 3.7　各グループの投稿元上位 10 件（2009 から 2010）

2009	2010
web(12.17%)	web(12.46%)
Echofon(7.73%)	twicca(8.80%)
twicca(6.35%)	TwitterforiPhone(7.57%)
TwitterforiPhone(6.29%)	ついっぷる/twipple(5.80%)
Tween(6.28%)	Echofon(5.06%)
SOICHA(4.16%)	SOICHA(4.97%)
Janetter(4.05%)	TwitterforAndroid(4.30%)
YoruFukurou(3.79%)	Janetter(3.96%)
TweetDeck(3.60%)	KeitaiWeb(3.60%)
ついっぷる/twipple(3.46%)	ついっぷる foriPhone(3.05%)

一方で，Twitter の利用開始時期が早いユーザが投稿数が多い結果となったことは，フォロー・フォロワー数 [6] との関連など，個々のユーザについて投稿活動を調査することでその理由を明らかにすることができると思われ，今後の課題と考えている．

表 3.8　各グループの投稿元上位 10 件（2011 から new comers）

2011	2012	new comers
TwitterforiPhone(19.35%)	TwitterforiPhone(24.11%)	TwitterforAndroid(20.29%)
TwitterforAndroid(11.17%)	TwitterforAndroid(18.73%)	TwitterforiPhone(14.63%)
web(10.84%)	web(9.34%)	web(14.04%)
KeitaiWeb(10.29%)	KeitaiWeb(9.01%)	KeitaiWeb(11.46%)
twicca(5.60%)	twittbot.net(4.88%)	twittbot.net(4.36%)
ついっぷる foriPhone(4.03%)	twicca(3.50%)	twicca(2.37%)
twittbot.net(3.38%)	ついっぷる foriPhone(2.86%)	ついっぷる/twipple(2.22%)
ついっぷる/twipple(3.35%)	Janetter(2.31%)	ついっぷる foriPhone(2.06%)
Echofon(2.52%)	TwippleforAndroid(2.28%)	MobileWeb(1.84%)
SOICHA(2.25%)	ついっぷる/twipple(2.18%)	TwippleforAndroid(1.68%)

3.5.2　投稿時刻に関する考察

3.4.2 節の投稿時刻に関する分析結果から，利用開始時期によらずにツイートの投稿数は昼食時と 17 時以降の夜間に増大していることがわかる．このことから，ユーザは仕事や授業の合間あるいは終了後の余裕のある時間にツイートを投稿していると思われる．

図 3.7 のリプライ率の推移において，夜間のリプライ率が大きくなったのは，その時間帯が多くのユーザにとって，比較的自由な時間帯にあたることや，先行調査 [12] で報告されているようにテレビ番組を視聴しながら意見の共有を目的とするユーザが存在することなどが原因であると思われる．

3.5.3　ツイートの投稿元に関する考察

3.4.3 節の投稿元に関する分析結果から，2006 年から 2008 年に利用を開始したグループでは，Twitter の公式の web の他に Mac OS 用のクライアントである YoruFukurou や Windows 用クライアントである Tween の投稿の割合が多い傾向を示しており，PC からの投稿が多くなされていることがわかる．続く 2009 年から 2010 年に利用を開始したグループでは，PC 用のクライアント以外にもスマートフォンからの投稿の割合が多くなっている．2011 年以降に利用を開始したグループでは，スマートフォンやフィーチャーフォンなどの携帯電話からの投稿が大きな値を示していることがわかる．該当するグループの投稿元の上位 4 位

の占める割合をみてみるとツイートの約半数が，Twitter for iPhone，Twitter for Android，Keitai Web といった携帯電話用のクライアントから投稿されている．これは，スマートフォンの普及や Twitter の認知度が上昇した影響で，携帯電話から利用するユーザが増加したことが原因の一つであると考えられる．

各グループの投稿元を俯瞰してみると，利用開始時期が遅くなるにつれて PC を中心に利用するユーザから携帯電話を中心に利用するユーザに変容していることが示唆される．加えて，ユーザは Twitter を使い始めた時に使用した投稿元を使い続ける恵子が強いことも明らかとなった．また，自宅では PC から投稿し，出先では携帯電話から投稿するというように複数の投稿元を使い分けるユーザが一定数存在すると思われ，今後より詳細な分析を行う必要がある．

3.6　まとめ

本章では，ユーザのプロファイルのうちユーザの利用開始時期に着目した分析を行っている．利用開始時期ごとにユーザのグループを作成し，それぞれのグループを比較した結果，Twitter の利用を開始した時期によってユーザの投稿活動が異なることが明らかになった．投稿数およびリプライ数に関する分析では，利用開始時期が遅いユーザにおいて，リプライの割合が高い傾向を示した．ツイートの投稿時刻に着目すると，利用開始時期が早いユーザは，日中の投稿の割合が多い結果となった．投稿元に関する分析では，利用開始時期が 2011 年以降のユーザの投稿は，スマートフォンやフィーチャーフォンなどの携帯電話からの割合が高いことが明らかとなった．

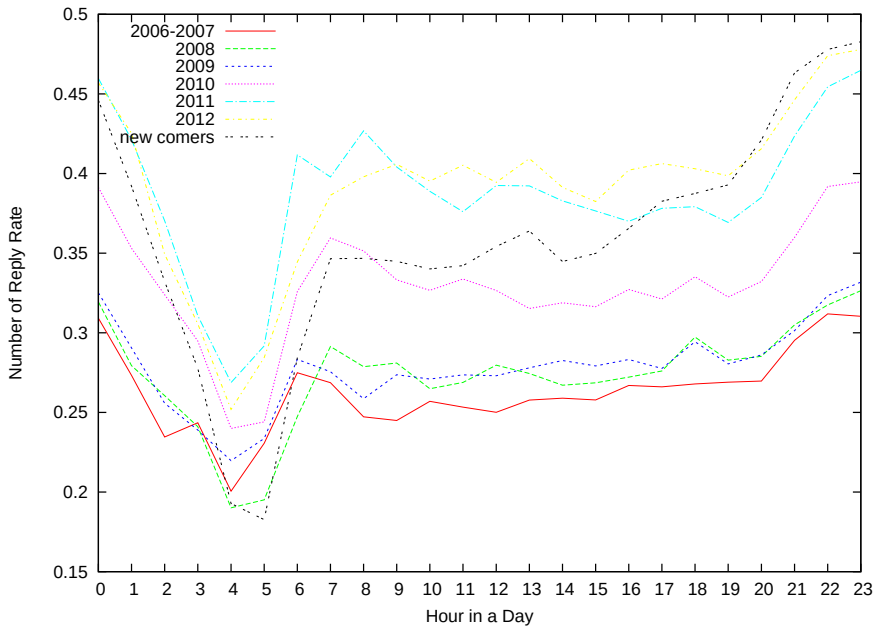


図 3.7　各時刻の投稿数に占めるリプライの割合

第 4 章

利用継続時間に着目した短期分析

4.1　はじめに

個々のユーザの投稿活動に着目すると，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る．そのため，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．加えて，Twitter を長期間使用し続けるユーザと短期間で Twitter の使用を辞めるユーザの投稿活動は異なると考えられる．

本章では，全てのユーザは単位時間に特定の投稿活動を示すクラスタに所属し，時間経過とともにクラスタ間を遷移するとみなし，投稿活動の時間縦断的分析手法を提案する．提案手法は，投稿活動の変化を表すクラスタ遷移系列の作成とクラスタ間の関係を表す状態遷移図の作成からなる．約 1 年間の日本語ツイートを対象に分析することで，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，利用を辞めるまでのライフサイクル解明を試みる．

本章の構成を以下に示す．4.2 節では分析手法を説明し，4.3 節，4.4 節で分析結果および考察を述べる．最後に，4.5 節でまとめを述べる．

4.2　投稿活動の遷移の分析手法

本章では，全てのユーザは単位時間に特定の投稿活動を示すクラスタに所属し，時間経過とともにクラスタ間を遷移するとみなし分析を行う．分析方法は，投稿活動の変化を表すクラス

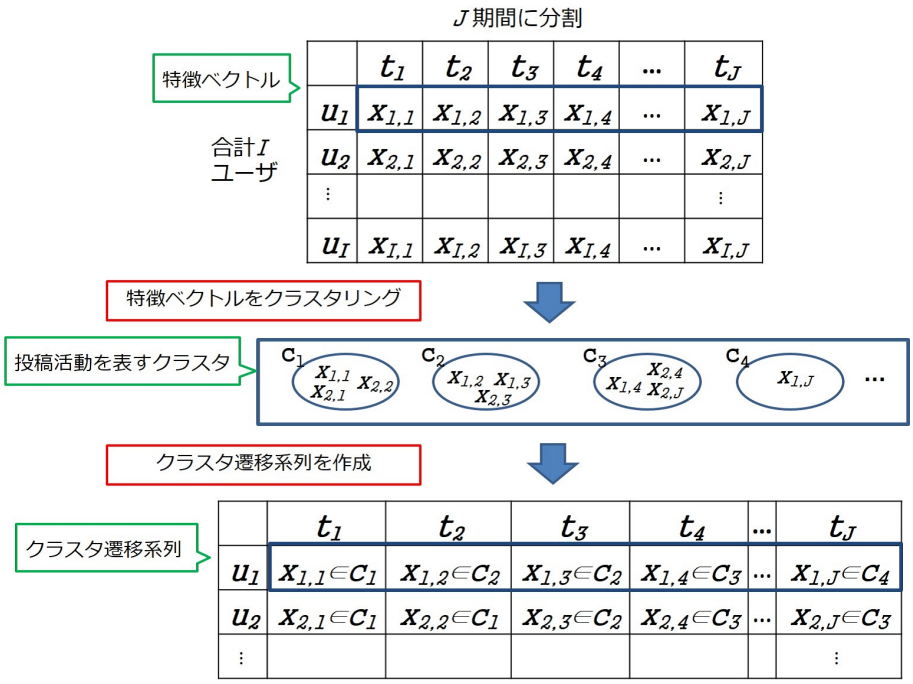


図 4.1　クラスタ遷移系列の作成方法

タ遷移系列の作成と，クラスタ間の関係を表す状態遷移図の作成からなる．クラスタ遷移系列の作成法について 4.2.1 節で，状態遷移図の作成法について 4.2.2 節で説明する．

4.2.1　クラスタ遷移系列の作成

投稿活動は投稿数や RT 数，リプライ数，ツイートの投稿時刻など多くの要素で構成される．あるユーザの投稿活動は時間的な変化が存在し，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考えられる．

本章においては，同一のユーザの特徴ベクトルであっても，異なる期間に抽出されたものは区別して扱う．したがって，特徴ベクトルは分析対象とする全ユーザごとに，分割期間数だけ抽出される．特徴ベクトルの総数は分析対象ユーザ数と分割期間数を乗算した数となる．全ユーザの特徴ベクトルをクラスタリングすることで，単位時間における特定の投稿活動を示すクラスタを作成できる．

ユーザの各期間の特徴ベクトルが所属するクラスタ番号を並べることで，クラスタ遷移系列を作成する．クラスタ遷移系列を用いて，投稿活動の時間遷移を分析する．

クラスタリングの概要を図 4.1 の上部に示す．クラスタリングの対象は，それぞれの分割期間ごとに抽出されたユーザの投稿活動を表す特徴ベクトルである．まず，特徴ベクトル集合 $X = \{x_{i,j} | user\ i(1 \leq i \leq I), timeperiod\ j(1 \leq j \leq J)\}$ を作成する，ここで，$x_{i,j}$ は，ユーザ $u_i(1 \leq i \leq I)$ が期間 $t_j(1 \leq j \leq J)$ に投稿したツイート集合 $D_{i,j}$ から作成された特徴ベク

トルである．

投稿活動は様々な特徴量から構成されると考えられるが，本章においては，以下の 11 種類の特徴量を使用する．

1 投稿数 (posts)
2 リプライ数 (replies)
3 異なりリプライユーザ数 (reply_user)
4 RT 数 (rt)
5 ツイートの平均文字数 (aver_char)
6-11 4 時間ごとのツイートの投稿数 (posts 0-3 to posts 20-23)

平均が 0，標準偏差が 1 となるように正規化した上記の特徴量を使用して，K-means 法により特徴ベクトルのクラスタリングを行う．クラスタリングの結果として，クラスタ集合 $C = \{c_l | 1 \leq l \leq K\}$ を得る．

クラスタ遷移系列の作成方法を図 4.1 の下部に示す．各ユーザごとに特徴ベクトル $x_{i,j}$ が所属するクラスタ番号を，時系列に従って並べることでユーザ u_i のクラスタ遷移系列を作成する．

4.2.2　状態遷移図の作成

クラスタ間の関係を表す状態遷移図を作成し，投稿活動の遷移を分析する．クラスタ間の状態遷移図を作成するために，全てのクラスタの組み合わせにおいて遷移確率を計算する．クラスタ c_l と c_m がクラスタリングの結果得られた場合，クラスタ c_l からクラスタ c_m への遷移確率 $P_{l,m}$ は次の式で計算される．

$$P_{l,m} = \frac{n_{l,m}}{\sum_{h=1}^{K} n_{l,h}}, \tag{4.1}$$

ここで $n_{l,m}$ は全ユーザの遷移系列においてクラスタ c_l の直後にクラスタ c_m へ遷移した頻度であり，K はクラスタ数である．クラスタ c_l からクラスタ c_m への遷移確率は，クラスタ c_l から遷移したクラスタの総数に占める，クラスタ c_m に遷移した回数の比率である．

図 4.2 に，3 ユーザの 4 期間のクラスタ遷移系列からの状態遷移図の作成の例を示す．以下の手順に従い状態遷移図を作成する．まず，クラスタ間の遷移回数を計算する．図 4.2 では，$n_{1,1}$, $n_{1,2}$ は 2 で，$n_{1,3}$ は 1 となっている．次に，遷移確率を計算する．図 4.2 では，$P_{1,1}$ は 0.5, $P_{1,2}$ と $P_{1,3}$ は 0.25 となる．最後にクラスタをノード，遷移の有無をエッジとして状態遷移図を作成する．

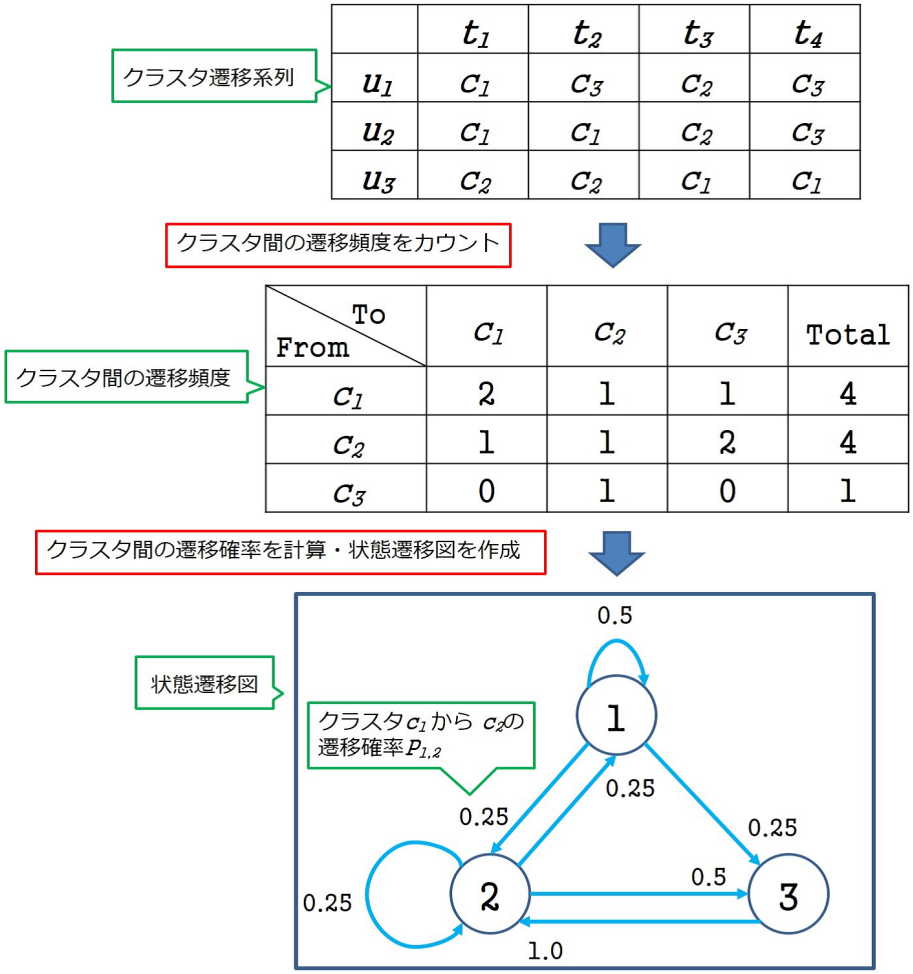


図 4.2　状態遷移図の作成例

4.3　短期分析の結果と評価

4.3.1　データセット

本節では，分析に使用するデータセットについて説明する．3.3.1 節で述べた方法で収集したツイートを母集団とし，特定の日時にアカウントを作成したユーザのツイートを分析用データセットとして用いる．データセットの概要を表 4.1 に示す．2011 年 11 月 16 日にアカウントを作成したユーザが，2011 年 11 月 16 日から 2012 年 11 月 13 日までの約 1 年間に投稿したツイートをデータセットとする．特徴ベクトル作成のために 7 日単位で 52 週に期間を分割した．

長期間 Twitter を使い続けるユーザと短期間で Twitter の利用を辞めるユーザの違いを分析するために，分析対象ユーザから利用継続時間が異なる 2 種類のグループ Long と Short を作成し比較する．2 グループの概要を 4.2 に示す．分析期間中の最も新しいツイートの投稿時

表 4.1　データセット概要

アカウント作成日	Nov 16, 2011
分析開始日	Nov 16, 2011
分析終了日	Nov 13, 2012
分割数	52
ユーザ数	8,417
ツイート数	2,802,317

表 4.2　グループ概要

グループ名		Long	Short
ユーザ数		1,000	1,000
ツイート数		1,281,338	38,509
利用継続時間 (days)	最大値	363.98	52.59
	最小値	350.57	7.01
	平均値	359.21	27.32
	標準偏差	3.64	13.12

刻と最も古いツイートの投稿時刻の差を利用継続時間とした．Long は利用継続時間の降順で並べた上位 1,000 ユーザであり，Short は利用継続時間が 7 日以上のユーザを利用継続時間の昇順で並べた上位 1,000 ユーザである．

4.3.2　投稿活動のクラスタリング結果

表 4.1 のデータセットを対象に，4.2.1 節の方法でクラスタリングを行った結果を示す．統計解析ツールの R [*1] を使用し, 特徴ベクトルを 20 クラスタに分割した．各クラスタに所属する特徴ベクトルの総数を表 4.3 に示す．表中の Long と Short のカラムはそれぞれのグループに所属するユーザの特徴ベクトル数である．クラスタの重心の特徴ベクトルの特徴量の総和の降順に，クラスタ番号を付与した．したがって，クラスタ 1 は重心の特徴ベクトルの特徴量の総和が最小のクラスタであり，クラスタ 20 は最大のクラスタである．

各クラスタの要素数に対して，グループ Long または Short に所属するユーザの比率が高いクラスタの特徴ベクトルを図 4.3 に示す．クラスタは以下の手順で選出した．まず，グループ Long または Short に所属する要素が 100 以上含まれるクラスタを選出する．グループ Long

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

表 4.3　クラスタの要素数

クラスタ番号	要素数	Long	Short
1	331,590	13,426	48,946
2	40,340	15,067	1,229
3	21,858	5,685	889
4	14,146	3,263	613
5	14,280	7,017	156
6	3,036	1,341	46
7	3,984	1,997	38
8	4,080	2,059	32
9	609	351	6
10	908	399	10
11	1,499	801	8
12	100	78	1
13	362	165	4
14	398	195	10
15	252	95	3
16	21	15	3
17	95	28	5
18	90	17	0
19	19	0	1
20	17	1	0

では，クラスタ 1，2，3，4，5，6，7，8，9，10，11，13，14 が条件を満たす．一方の，グループ Short では，クラスタ 1，2，3，4，5 が該当する．次に，選出したクラスタの中からクラスタの要素数に対する，それぞれのグループの要素数の比率が高い上位 5 クラスタをそれぞれのグループを代表するクラスタとする．例えば，クラスタ 9 の場合は，クラスタ 9 の要素数 609 のうち，グループ Long に所属する要素数は 351 なので比率は 0.57 となる．グループ Long では，クラスタ 9，11，8，7，5 が，グループ Short では，クラスタ 1，4，3，2，5 が選出される．図 4.3 から，グループ Long の要素の比率が高いクラスタ 5，8，7，11 はリプライ数が大きく，クラスタ 9 は RT 数が大きいことがわかる．一方でグループ Short の要素の比率が高いクラスタ 2，3，4 はリプライ数，RT 数ともに小さい．また，クラスタ 1 は重心の特徴量の総和が最小のクラスタであり，ユーザがツイートを投稿していない状態を表している．

4.3.3　状態遷移図

全ユーザのクラスタ遷移系列から作成した状態遷移図を図 4.4 に示す．図中で，ノードはクラスタで，ノードに付与された数字はクラスタ番号である．クラスタ c_l と c_m の間のエッジに付与された値は，遷移確率 $P_{l,m}$ である．なお，図中では遷移確率が 0.1 以上のエッジのみ図示している．上述の理由から，あるノードからの遷移確率の総和は 1.0 とはならない．図 4.4 ではクラスタ 1 は他のクラスタへ遷移するエッジを持たないことがわかる．

グループ Long と Short に所属するユーザの状態遷移図を図 4.5 と図 4.6 に示す．図 4.5 では，図 4.4 と異なり，クラスタ 2 からクラスタ 5，クラスタ 1 からクラスタ 2 へ遷移するエッジが存在している．図 4.6 においては，クラスタ 2 からクラスタ 1 への遷移確率が大きく，多くのクラスタにクラスタ 1 へ遷移するエッジが存在している．

4.4　考察

図 4.4 の，全ユーザの状態遷移図はクラスタ 1 に向かい収束する形状となっている．クラスタ 1 はツイートを投稿しない状態を表すクラスタであり，このことから，一度ツイートの投稿を辞めたユーザは投稿を再開しにくいことがうかがえる．

図 4.5 から，グループ Long に所属するユーザの状態遷移図の特徴として，クラスタ 1 からクラスタ 2，クラスタ 2 からクラスタ 5 へ遷移するエッジが挙げられる．クラスタ 1 からクラスタ 2 へ遷移するエッジは，一度投稿を辞めたユーザが投稿を再開することを意味しており，状態遷移図が，長期間 Twitter の利用を継続するユーザの特徴を表現できていると考えられる．また，クラスタ 2 からクラスタ 5 へ遷移するエッジが存在するのは図 4.5 のみとなっている．クラスタ 5 は重心の特徴ベクトルのリプライ数がクラスタ 2 より大きいクラスタである．加えて，図 4.3 においてグループ Long の要素の比率が大きいクラスタにおいては，RT 数やリプライ数が大きかった．このことから，長期間 Twitter の利用を継続するユーザを特徴付ける要因の 1 つとして，リプライや RT の使用の頻度が挙げられる．

図 4.6 において，多くのクラスタにおいてクラスタ 1 へ遷移するエッジが存在している．これは，グループ Short のユーザが Twitter の利用を休止しする傾向が大きいことを表している．

グループ Long と Short を比較した結果，リプライや RT を用いた他のユーザとのコミュニケーションの有無が，長期間 Twitter を利用するユーザと短期間で利用を辞めるユーザの違いの 1 つとして挙げられる．

4.5　まとめ

本章では，ユーザのプロファイルのうちマイクロブログの利用継続時間に着目した分析を行った．全てのユーザは単位時間に特定の投稿活動を示すクラスタに所属し，時間経過とともにクラスタ間を遷移するとモデル化し，投稿活動の時間縦断的分析手法を提案した．提案した分析方法を用いて，長期間 Twitter の利用を継続するユーザと短期間で利用を辞めるユーザの違いを比較した．

分析の結果，長期間利用を継続するユーザと短期間で利用を休止するユーザでは，投稿活動の時間遷移の過程が異なっていることが明らかとなった．また，リプライや RT を用いた他のユーザとのコミュニケーションの有無が，長期間 Twitter を利用するユーザと短期間で利用を辞めるユーザの違いの 1 つとして示唆された．

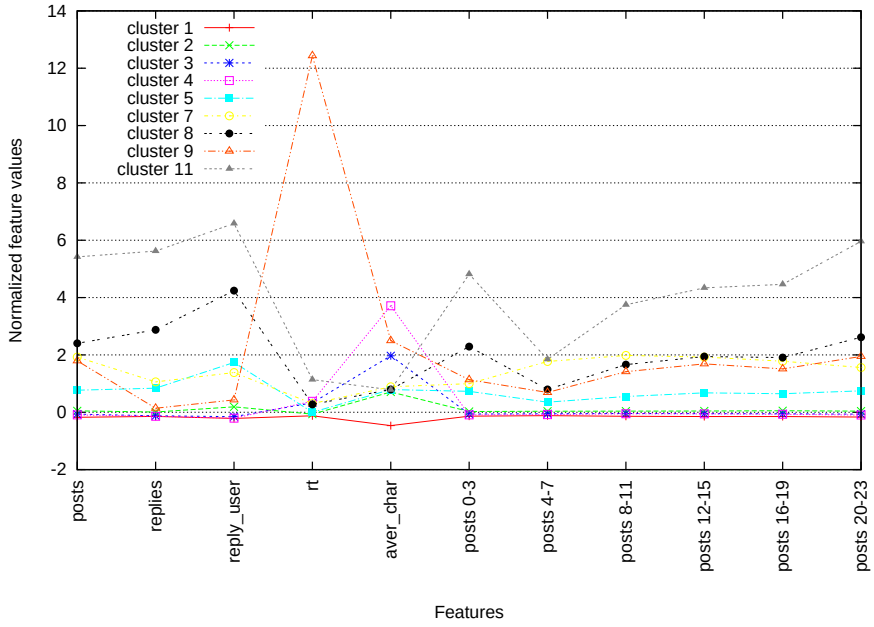


図 4.3　2 グループの要素数の比率が大きいクラスタの重心の特徴ベクトル

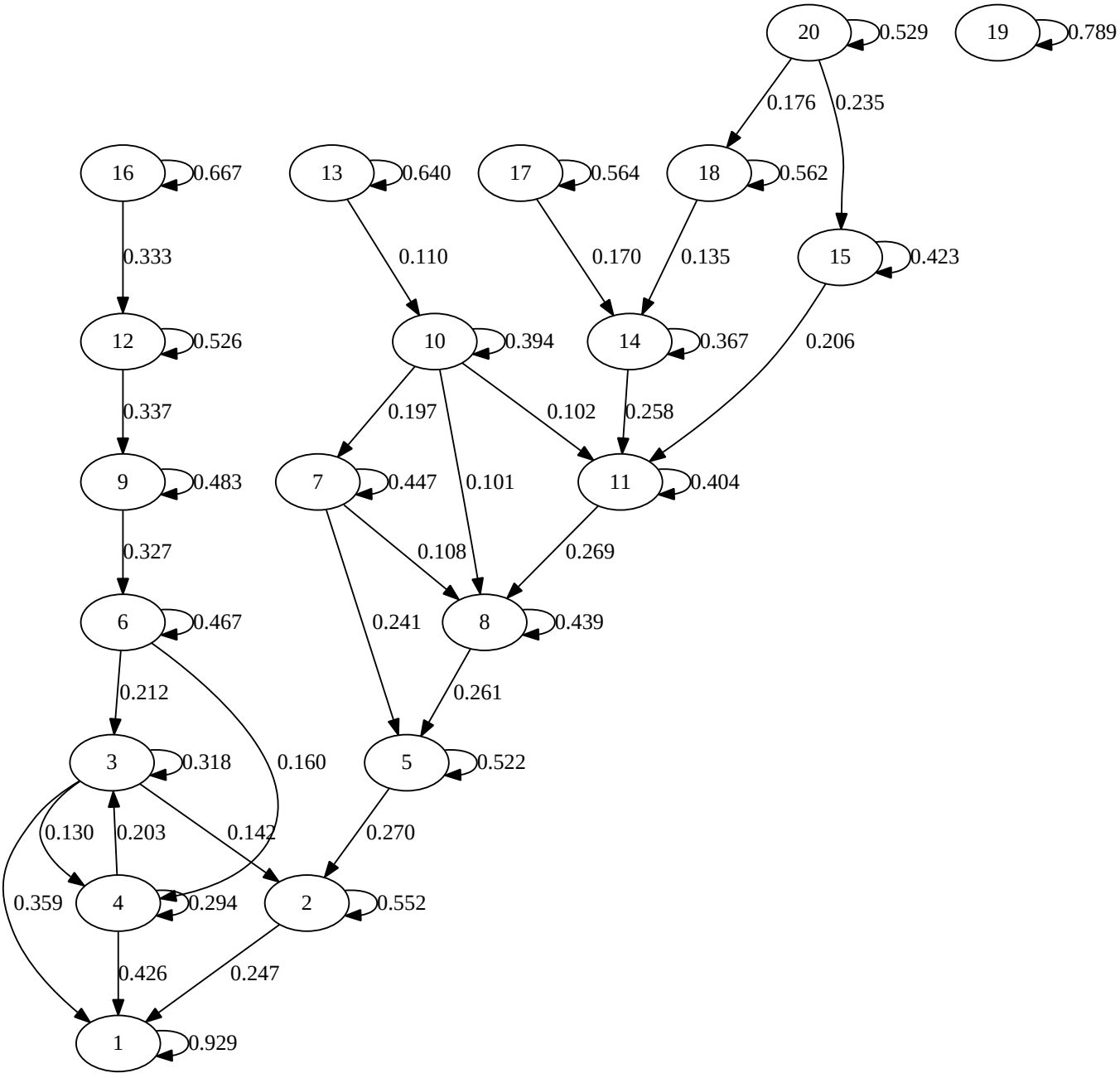


図 4.4　全ユーザの状態遷移図

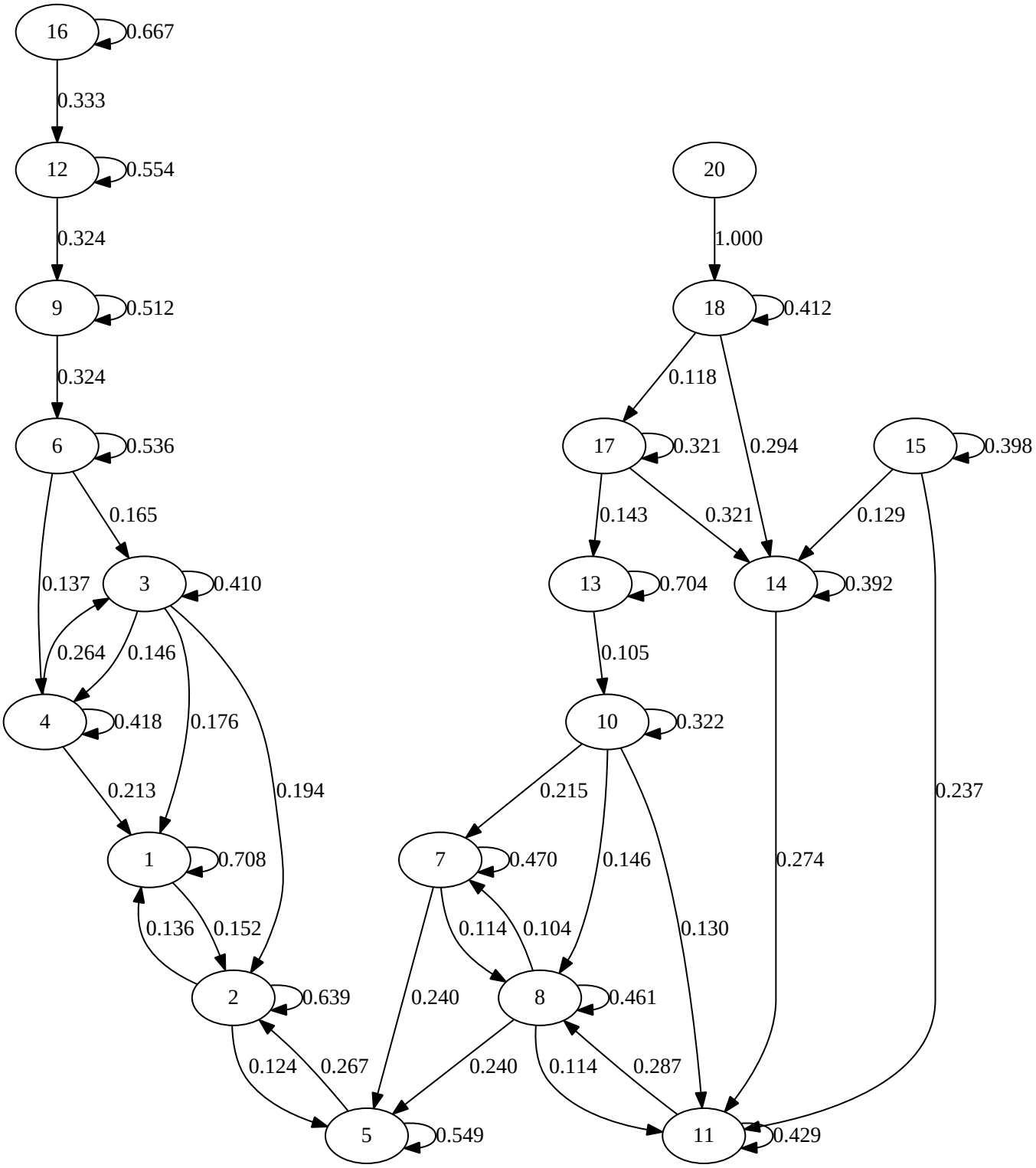


図 4.5　グループ Long の状態遷移図

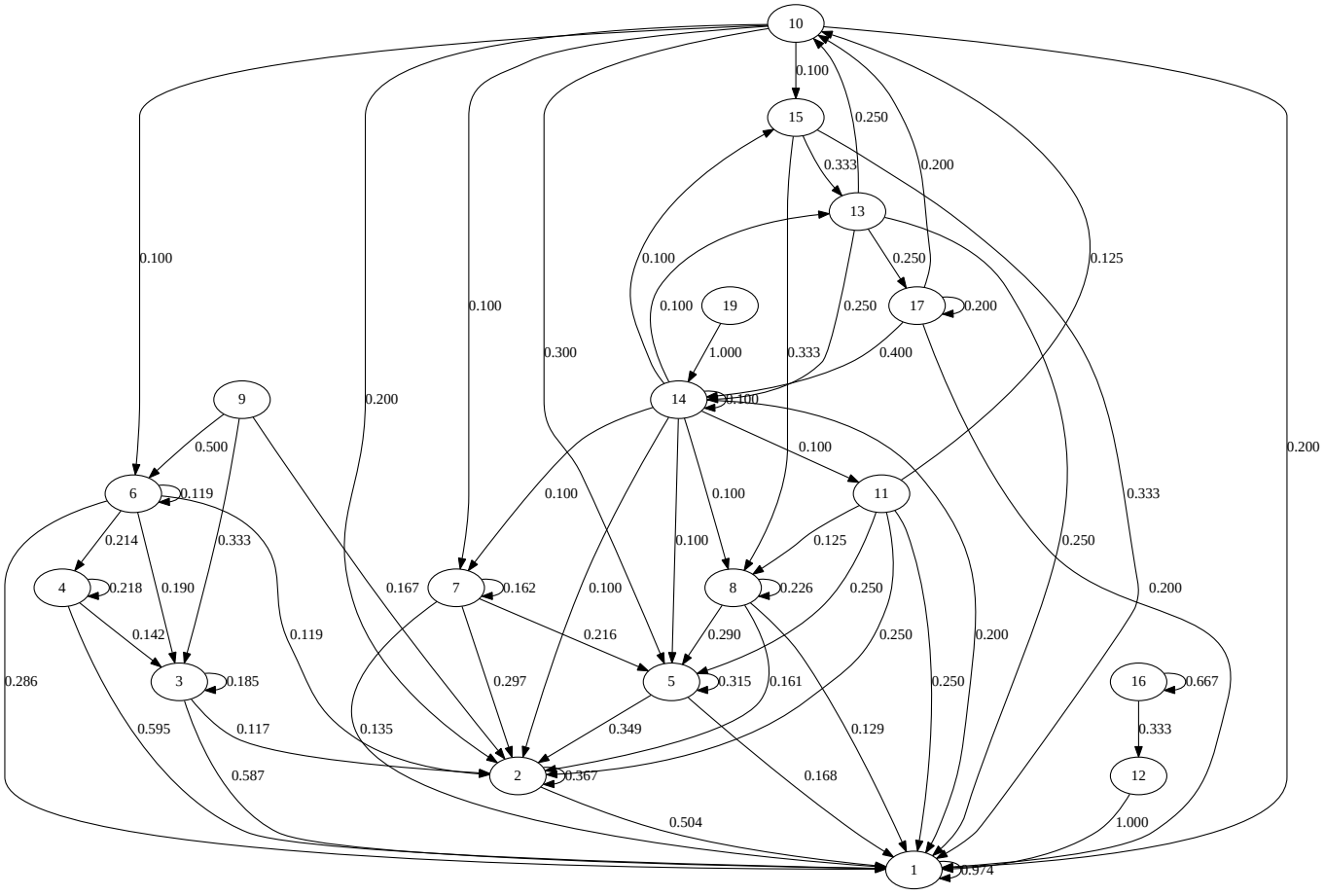


図 4.6　グループ Short の状態遷移図

第 5 章

考察

本章では，3 章と 4 章で得られた分析結果について考察を行う．

3 章では，ユーザのプロファイルのうち利用開始時期に着目した分析を行った．利用開始時期ごとにユーザのグループを作成し，それぞれのグループに対し分析を行うことで，Twitter の利用を開始した時期によってユーザの投稿活動が異なることが確認できた．利用開始時期が早いユーザと遅いユーザの間の顕著な差異として，投稿に用いる端末や投稿数に占めるリプライの割合が異なることが明らかとなった．

4 章では，ユーザのプロファイルのうちマイクロブログの利用継続時間に着目した分析を行った．全てのユーザは単位時間に特定の投稿活動を示すクラスタに所属し，時間経過とともにクラスタ間を遷移するとみなす方法を用いて，利用継続時間が長いユーザと短いユーザの比較を行った．長期間利用を継続するユーザの比率が多いクラスタでは，単位時間当たりのリプライや RT 数が多い結果となり，他のユーザとのコミュニケーションの有無が，長期間 Twitter を利用するユーザと短期間で利用を辞めるユーザの違いの 1 つとして示唆された．

本研究で用いた 2 つのプロファイル，利用開始時期と利用継続時間については，投稿活動との関連が明らかとなった．ここから，投稿活動に基づいたユーザのプロファイリングの可能性が示せたと考える．

第 6 章

結論

6.1　まとめ

本研究ではプロファイルの異なるマイクロブログユーザ群，あるいは個々のユーザの特徴を投稿活動を通して明らかにするために，利用開始時期と利用継続時間に着目した分析を行った．利用開始時期に着目した長期分析では，利用開始時期ごとにユーザのグループを作成し，それぞれのグループを比較することで，Twitter の利用を開始した時期によってユーザの投稿活動が異なることを明らかにした．マイクロブログの利用継続時間に着目した短期分析では，個々のユーザの投稿活動の変化を状態遷移図を用いて分析する手法を提案した．提案手法を利用継続時間が長いユーザと短いユーザに適用し比較した．その結果，長期間利用を継続するユーザと短期間で利用を休止するユーザでは，投稿活動の遷移過程が異なる傾向を示した．また，長期間利用を継続するユーザの特徴としてリプライや RT 数が多いことが挙げられ，他のユーザとのコミュニケーションの有無が，長期間 Twitter を利用するユーザと短期間で利用を辞めるユーザの違いの 1 つとして明らかになった．以上の分析によって，投稿活動に基づくマイクロブログユーザのプロファイリングの有効性を示すことができた．

6.2　今後の課題

今後の課題として，マイクロブログの利用継続時間の推定や，投稿活動の遷移の時系列分析における分割期間の粒度の変更が挙げられる．また，本研究では日本国内でツイートを投稿したユーザを対象に分析を行ったが，他言語のユーザに対しても分析を行うことが挙げられる．

謝辞

本論文は，筆者が筑波大学大学院図書館情報メディア研究科博士前期課程に在籍中の研究成果をまとめたものである．同研究科の佐藤哲司教授には主指導教員として，卒業研究から 3 年間にわたりご指導をいただいたこと，謹んで感謝申し上げます．副指導教員として，研究の節目で適切なご助言をいただいた，関洋平助教にも感謝の意を表し，お礼申し上げます．また，池内淳准教授には，学群時はクラス担任として，大学院では共著者として丁寧にご指導いただき，大変感謝しております．

研究を進めるにあたり，実験環境や研究室運営について尽力し，研究についても常に適切な助言をして頂いた，研究室 OB である，法政大学マイクロナノテクノロジー研究センターの島田諭さんにも大変感謝しております．

研究室の同期として共に 3 年間助けあいながら，研究を進めてきた山本修平くん，そして関研究室所属の大山鉄郎くん，堂前友貴さんにも大変感謝しております．ありがとうございました．本研究を最後までやり遂げることができたのは，先生方，先輩方，後輩たち，同期の皆様のおかげです．本当にありがとうございました．

参考文献

[1] 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.

[2] 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.

[3] Gideon Dror, Dan Pelleg, Oleg Rokhlenko, and Idan Szpektor. 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.

[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] Lichan Hong, Gregorio Convertino, and Ed H. Chi. Language matters in twitter: A large scale study. *Proceedings of the 5th International Conference on Weblogs and Social Media (ICWSM-11)*, pp. 518–521.

[6] Bernardo A. Huberman, Daniel M. Romero, and Fang Wu. Social networks that matter: Twitter under the microscope. *CoRR, Vol.abs/0812.1045*, 2008.

[7] 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.

[8] 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.

[9] 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）.

[10] 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.

[11] 株式会社トライバルメディアハウス, 株式会社クロス・マーケティング（編）. ソーシャルメディア白書 2012. 翔泳社, 2012.

[12] 渋谷明子, 志岐裕子, 李光鎬. Sns 利用者のコミュニケーションとテレビ視聴： ウェブ・モニター調査 (2011 年 2 月) の報告 (2) (特集 ネット時代のテレビの役割). メディア・コミュニケーション： 慶応義塾大学メディア・コミュニケーション研究所紀要, No. 62, pp. 57–78, mar 2012.

[13] 総務省情報通信国際戦略局情報通信経済室. 次世代ＩＣＴ社会の実現がもたらす可能性に関する調査研究. http://www.soumu.go.jp/johotsusintokei/linkdata/h23_05_houkoku.pdf（参照 2012-10-12）.

[14] 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.

[15] 富士通総研. Twitter（ツイッター）利用状況調査. http://jp.fujitsu.com/group/fri/report/cyber/research/twitter/（参照 2012-10-12）.

発表論文

国際会議論文

- Yutaro Yamaguchi, Shuhei Yamamoto, and Tetsuji Satoh. Behavior Analysis of Microblog Users Based on Transitions in Posting Activities, 15th International Conference. Information Integration and Web-based Applications & Services(iiWAS2013), pp. 63-67, 2013.

国内会議論文

- 山口 裕太郎，山本 修平，佐藤 哲司. 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, ARG, Web インテリジェンスとインタラクション研究会 (ARG SIG-WI2), pp.7 - 12, 2013.
- 山口 裕太郎, 山本 修平, 佐藤 哲司. 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, 情報処理学会, マルチメディア，分散，協調とモバイルシンポジウム (DICOMO2013), 1C-3, pp. 72 - 79, 2013.
- 山口 裕太郎, 山本 修平, 島田 諭, 佐藤 哲司. マイクロブログにおける利用目的の変容過程に着目したユーザプロファイル分析手法の提案, 情報処理学会, 情報アクセスシンポジウム 2012(IAS2012), pp. 8-14, 2012.
- 山口 裕太郎, 山本 修平, 水沼 友宏, 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, 情報社会学会, 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.
- 山口 裕太郎, 島田 諭, 佐藤 哲司: 人物の呼称に基づくマイクロブログ記事における話題の時間的推移に関する一考察, 情報処理学会, マルチメディア，分散，協調とモバイルシンポジウム (DICOMO2012), 8H-34, pp. 2279 - 2286, 2012.