Commentary

大規模ソーシャル・ネットワークはどのように成長したか?

――Twiitter ネットワークの収集から解析まで――

渡部　優（東京工業大学大学院情報理工学研究科修士課程／JST CREST）
鈴村豊太郎（東京工業大学大学院情報理工学研究科客員准教授／JST CREST）

1　はじめに

近年，Twitter, Facebook, MySpace, LinkedIn といったソーシャルネットワークサービス（SNS）が急速に普及し，私たちの生活に深く浸透してきた。これらのサービスは人と人をつなげるコミュニケーションツールとして大きな役割を果たすようになり，それと同時にそのコミュニティのあり方も多様化してきた。たとえば，Facebook では，実名を使ったアカウント登録が行われ，実生活における友人・知人とのネットワークが作られる。Twitter では，ハンドルネームを使ったアカウント登録が行われ，ユーザの興味や関心のある人とのネットワークが作られる。また，MySpace では，音楽を中心としたサービスが展開され，同じ趣味をもった人々の間でネットワークが作られる。

このようなネットワークを解析することで，多様なコミュニティがそれぞれどのような性質や構造をもっているかが理解できる。しかし，最近では，ユーザの増加に伴うネットワークの成長とデータの巨大化により，データの取得や解析に大きなコストがかかり，手軽に扱うことができなくなってきている。私たちは，このような巨大化するデータに対して，スパコンなどの高性能計算機を用いることで，大規模なネットワークの解析を行っている。

本稿では，2012 年 10 月，ユーザ数が 4.7 億人に達する Twitter の実ネットワークのデータ収集から，東京工業大学のスーパコンピュータ（以下，スパコン）TSUBAME 2.0 を用いた解析までを紹介する。また，ネットワーク成長という観点から，Twitter がどのように成長してきたのか，Facebook ネットワークなどの比較も織り交ぜて紹介していく。

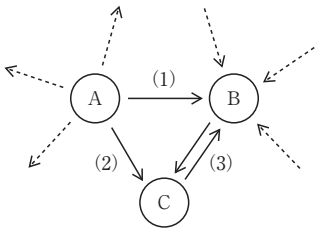
2　大規模ソーシャル・ネットワークとしての Twitter

サービス概要

ここでは，Twitter のサービスとそこで構築されるネットワークについて簡単に説明する。Twitter とは，140 文字以内のショートメッセージ（“ツイート”）を投稿できるマイクロブログサービスである。このツイートは投稿者のタイムラインに保存され，ユーザはこの投稿者のタイムラインをいつでも閲覧することができる。また，タレントやアーティストなどの気に入った投稿者がいれば，その投稿者を“フォロー”することによって，自身のタイムラインにそのツイートを表示することができるようになる。Twitter ではこのフォロー機能によって，**図 1** のように，たくさんのユーザをフォロー関係で結んだ巨大なネットワークが作られていく。なお，本稿では便宜上，フォロー関係の定義として，A が B をフォローしているとき，A を B の“フォロワー”，B を A の“フレンド”と呼ぶことにする。

大規模な Twitter ネットワークの収集

Twitter では，クライアント用に Twitter REST API（https://dev.twitter.com/docs/api）が提供されている。この REST API を利用すると，Twitter ユーザのアカウント情報やフォロー関係などを収集することができる。そこで私たちは，2012 年 8 月から 10 月にかけて REST API（v1.0）



(1) A が B をフォローしている
(2) A が C をフォローしている
(3) B と C は互いにフォローしている

図 1　フォロー関係による Twitter のネットワーク

を用いた自動収集（クロール）を行い，ユーザのアカウント情報とユーザのフォロワー・フレンド情報の 2 種類のデータを収集した。アカウント情報は，ユーザ ID，ユーザネーム，自己紹介文，アカウント作成日，地域，タイムゾーンなどのプロフィール情報である。フォロワー・フレンド情報とは，ユーザのフォロワーとフレンドの ID のリストである。これらのデータを集めるにあたり，私たちは以下の手順でクロールを行った。

まず初めに，レディー・ガガ（@lady_gaga）やオバマ大統領（@BarackObama）といったフォロワー数の最も多いトップ 1,000 のユーザをターゲットにする。このトップ 1,000 のユーザに対し，REST API を用いることでアカウント情報とフォロワー・フレンド情報を取得する。そして，フォロワー・フレンド情報を基に，トップ 1,000 のフォロワーすべてを新しいターゲットとし，アカウント情報とフォロワー・フレンド情報を取得する。以下，この作業を繰り返すことで，フォロー関係をたどってたくさんのユーザ情報を収集する。

私たちはこの作業を 28 回繰り返したところで，新しく取得できるユーザ数が 100 以下となったため，3 ヵ月に及ぶクロールを終了した。こうして，約 4.7 億ユーザのアカウント情報と 300 億のフォロー関係から構成されるフォロワー・フレンドネットワークを収集することに成功した。過去 2009 年にも，同様の手法（Kwak et al., 2009）で約 4,200 万ユーザからなる Twitter のネットワークが収集されているが，本稿におけるネットワークの規模は過去に例がないほど，大きなものとなっている。なお，データサイズとしては，アカウント情報とフォロワー・フレンド情報はそれぞれ圧縮状態で 91 GB と 231 GB に達する。

なお，本稿では Twitter Rest API を簡単に利用するため，Java 実装の Twitter4J（http://twitter4j.org/ja/index.html）というライブラリを使用している。Java 以外にも Python などでパッケージされたものも公開されているので，関心のある方は是非参考にしていただきたい。

3　Twitter の成長

私たちは第 2 節で収集した 4.7 億人のアカウント情報を利用し，ユーザ数の観点から Twitter のネットワークがどのように成長してきたのか調査した。この調査では，東京工業大学のスパコン TSUBAME 2.0 上で Apache Hadoop（http://hadoop.apache.org/）という大規模データの分散処理に特化した処理基盤を用いることで高速に集計を行った。ユーザのアカウント情報（91 GB）を入力データとし，計算機 4 台による並列計算でおよそ 30 分をかけて結果を算出した。

まず，Twitter のサービスが開始された 2006 年 6 月から 12 年 9 月までのユーザ数の変化を追ってみる。**図 2** は，ユーザのアカウント情報に含まれるアカウント作成日をもとに，ユーザ増加数を月別に集計したものを示している。

この図をみると，2008 年末までは大きな変化がみられないが，09 年の 1 月以降から，急激にユーザ数が伸びていることがわかる。しかし，この増加数の伸びは安定したものではなく，2009 年，10 年，11 年，12 年にやや波があるようにみえる。私たちはこの変化について，世界各地で Twitter の流行に時間差があったのではないかと考え，さらにユーザのアカウント情報（“タイムゾーン”）から地域別に分類を行い，変化を追った（図 3）[1]。

図 3 をみると，まず 2008 年から北米（アメリカ，カナダ）で徐々にユーザが増え始め，2009 年初めに爆発的に増加している。北米に続いて，ヨーロッパ・中南米，やや遅れて，アジアでユーザ数が伸び始めていることがわかる。

このグラフの特徴的な変化として，中南米でユーザ増加のピークが 2009 年，10 年，11 年と 3 回にわたっていることが確認できる。これは，エクアドル，ブラジル，チリでのユーザ増減がほぼ同

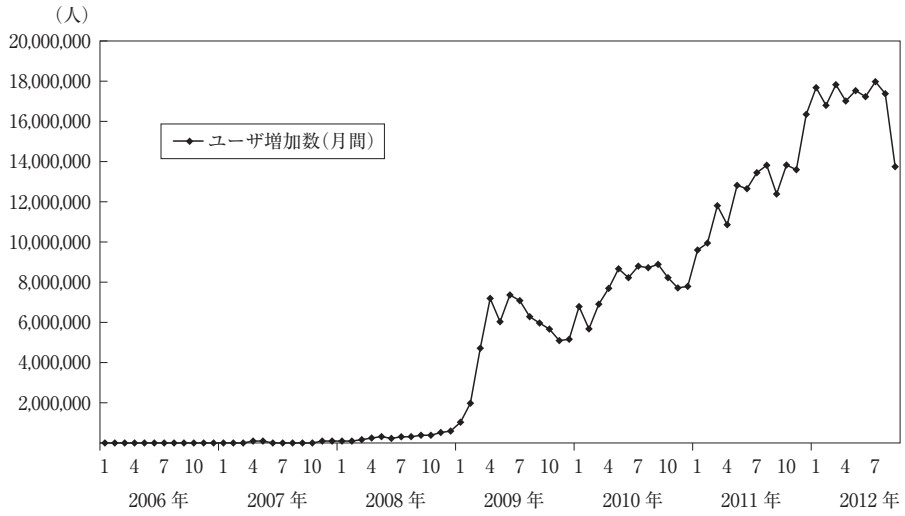


図２　Twitter のユーザ増加数の推移

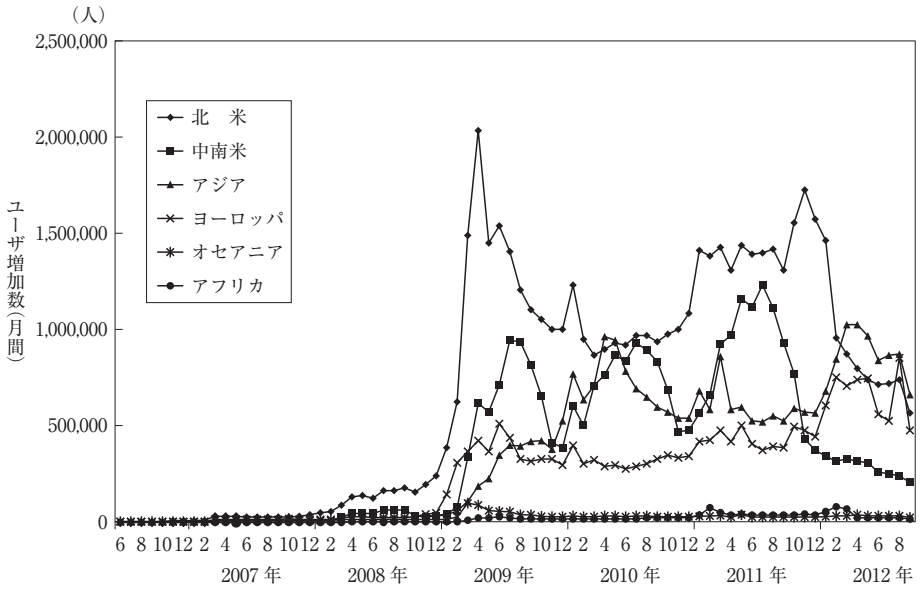


図３　ユーザ増加の地域別分類

じような波を描き，重なり合って現れたものである。また，アジアでも，2010 年 4 月，11 年 3 月，12 年 3 月と 3 回のピークがあるが，前 2 回のピークは日本におけるユーザの急激な増加が強く現れたもので，3 回目のピークはタイとイラクでのユーザ数の伸びが重なり合って現れたものとなっている。ヨーロッパにおいては，イギリスで 2009 年初めにユーザが一時的に大きく増加した後，オランダ・スペイン・ドイツ・ギリシャなどヨーロッパ各国でユーザが増え始めたため，ヨーロッパ全体として比較的安定した推移をみせている。2012 年になると，2 月にオランダ，4 月にギリシャ，8 月に再びイギリスでユーザが急増した。8 月のイギリスのユーザ増加は，2009 年ピーク時の 2 倍近くにもなる。

このように，ユーザ増加の推移を地域別に分類したことで，完全ではないものの，各地域・国で Twitter がどのように流行してきたのかが判明し

た。中南米の例では一部の国に相関があるような振る舞いも観測されるなど，世界各国でのTwitterの流行時期が推移していく様子がみてとれる。このことから，2006年のサービス開始から12年にかけて，Twitterのネットワークの構造が大きく変化してきたと考えられる。以降の節では，このようなネットワーク構造の変化を捉えるべく，2009年と12年のTwitterネットワークの性質を比較していく。

4 ネットワーク解析

隔たり

ネットワークの性質を調べる指標として，人と人の間の距離を測る“隔たり”がある。隔たりに関する研究としては，1967年に心理学者のスタンレー・ミルグラムらが行ったスモールワールド実験（Milgram et al., 1969）が有名である。彼らの実験では，無作為に選んだ人からある特定の人への手紙を人づてに届けてもらい，手紙が届くまでに何人の人を介したかを調べることで，ネットワークの隔たりを検証した。彼らはこの実験から，手紙が届くまでに平均5.8人の知人を仲介したという結果を得たが，現在ではこの実験結果を裏づけとし，世界の人々は6人の知人を介してつながっている“6次の隔たり”仮説がSNS上のネットワーク解析における下地となっている。

また最近では，Facebookユーザ7億人を対象とした解析（Backstrom et al., 2012）において，4.74（仲介する知人は3.74）の隔たりが発表され大きな話題となった。Facebookにおけるネットワークは，実名登録というサービスの形態上，実世界の知人のネットワークにかなり近いものになっていると考えられが，“6次の隔たり”仮説よりもはるかに値が小さい。

本稿では，Twitterのネットワークを対象に解析を行っているが，ではTwitterのネットワークとFacebookのネットワークではどのような違いがあるのか，非常に興味深い。Twitterの場合，ハンドルネームによる登録で，興味のある人を誰でもフォローできるため，Facebookとはネットワークの性質が大きく異なってくると考えられる。そこで私たちは，2009年と12年のTwitterネットワークの隔たりを計算し，Facebookの隔たりと比較・検証した。

1. 実験環境

隔たりの計算には，Facebookの解析（同上）でも用いられたJava実装のネットワーク解析ライブラリWebGraph（Boldi et al., 2011; http://webgraph.di.unimi.it/）を使用している。このライブラリを用いることで，隔たりの近似値を高速に求められるほか，ネットワークの直径の下限値も同時に求めることができる。なお，直径とは，隔たりが任意の2点間の最短距離の平均値であるのに対し，任意の2点間の最短距離の最大値となる。

また，本実験におけるTwitterのデータは4.7億ユーザと非常に規模が大きいため，東京工業大学のスパコンTSUBAME 2.0のなかで最も大きい512 GBのメモリを載せた計算機1台を使用した。入力データは，2009年時点の4100万ユーザ・15億フォローからなるネットワークと，12年時点の4.7億ユーザ・300億フォローからなるネットワークデータである。

2. 結果と考察

私たちは，2009年と12年のTwitterネットワークそれぞれにおいて，WebGraphよる計算試行を4回行い，その平均値をとることで隔たりを算出した。1回の試行には最大で14時間を要したが，これによって，2009年のネットワークで4.50，12年のネットワークで4.59という結果が得られた。2009年と12年で比較をすれば，12年ではわずかに人と人の間の距離が遠くなったといえる。しかしながら，Facebookの解析結果が4.74という“6次の隔たり”よりも小さい値であったのに対し，Twitterが4.59というさらに小さい値となったのは驚きである。この理由としては，Facebookがあくまで実世界の知人関係を主としてネットワークであるのに対し，Twitterは実世界では関係をもてないようなアーティストやタレントといったリンク数の高い有名人とフォロー関係をもつことができるため，そのような有名人を仲介として，人と人の間の距離がさらに短くなったと推測される。

また，直径については，2009年で25，12年に71という結果を得た。3年間で直径が大きくなった理由としては，第3節で述べたように，2009年の時点でTwitterが北米を中心としたサービス

であったのに対し，12年には世界各地でTwitterの利用者が増え，ユーザの関心や言語に壁ができたことが理由として考えられる。

相互性

ネットワークの相互性とは有向グラフ特有の指標で，相互フォローの関係がネットワーク中にどの程度存在しているのかを示すものである。Flickr（Cha et al., 2009）やYahoo! 360（Kumar et al., 2006）におけるネットワークの相互性は，それぞれ68%，84%と報告されており，友人やコミュニティとしてのSNSの性質が大きく表れているが，Twitterでは実際はどのような性質を示すのか。

今回，私たちが収集したTwitterの実ネットワークを用いて相互性の解析を行ったところ，2009年のネットワークで22.1%の相互性（Kwak et al., 2xxx）であったのに対し，12年のネットワークでは19.5%という結果を得た。FlickrやYahoo! 360と比べTwitterの相互性はかなり低いことがわかるが，これはTwitterがコミュニティベースのSNSというよりも，ツイートによる情報の流通を主としたマイクロブログサービスとしてのSNSであるからと考えられる。また，2009年から12年にかけて相互性が低下した点については，Twitterが世界的に広まったことが要因の1つだと私たちは推測している。第3節で示したように，2009年時点では北米を中心としたサービスであったが，時間経過とともに様々な地域のユーザに使われるようになったことで，ユーザの間にある関心や習慣，言語といったギャップがいっそう大きくなったことによると考えられる。

5 まとめ

本稿では，Twitter Rest APIを用いたデータ収集とスパコン上での大規模なネットワーク解析により，Twitterが世界的にどのように広まっていったのか，また，ソーシャルネットワークとしての性質として隔たりや相互性がどのように変化したのかを紹介した。

しかしながら，私たち計算機科学者にとっては，このような大規模データを解析することができても，社会学的にどのような価値に結びつけられるのか十分に理解できていない側面がある。その一方，社会学者にとっては，大規模データの扱いは敷居が高く，独自に解析を進めるのは難しいのではないだろうか。

海外ではこのような課題に対し，日々巨大化するデータに対応するため，社会学者と計算機科学者の協調研究が盛んに行われるようになってきている。今後，国内においても，このような動きを活発にしていくことが求められると私たちは考えている。

注

・1　タイムゾーンはデフォルトで“null”や“Hawaii”に設定されるため，地域の分類に利用できないアカウントが多い。ここでは設定が有効な1.3億のユーザから算出している。

文献

Backstrom, L. et al., 2012, “Four Degrees of Separation,” in *ACM Web Science 2012: Conference Proceedings*, ACM Press, 45–54.

Boldi, P. et al., 2011, “HyperANF: Approximating the neighbourhood function of very large graphs on a budget,” in *Proceedings of the 20th International Conference on World Wide Web*, ACM, 625–34.

Cha, M. et al., 2009 “Ameasurement-drivenanalysis of information propagation in the Flickr social network,” in *Proceedings of the 18th International Conference on World Wide Web*, ACM.

Kumar, R. et al., 2006, “Structure and evolution of onlinesocial networks,” in *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM.

Kwak, H. et al., 2009, “What is Twitter, a social network or a news media?,” in *Proceedings of the 19th International Conference on World Wide Web*, ACM.

Milgram, S. et al., 1969, “An experimental study of the small world problem,” *Sociometry*, 32（4）: 425–43.

参照URL

WebGraph（http://webgraph.di.unimi.it/）.

Apache Hadoop（http://hadoop.apache.org/）.

Twitter REST API（https://dev.twitter.com/docs/api）.

Twitter4J（http://twitter4j.org/ja/index.html）.