社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

[招待講演] 通信行動データを利用したコミュニティ計測に関する研究の動向

津川　　翔†

† 筑波大学 システム情報系
〒 305–8573 茨城県つくば市天王台 1–1–1
E-mail: †s-tugawa@cs.tsukuba.ac.jp

あらまし　ネットワーク利用者の行動に関する膨大かつ粒度の細かいデータが入手可能になっている。ソーシャルメディアにおける投稿や、電子商取引サイトにおけるレビュー文書、Web 検索エンジンにおけるトレンドなど、ネットワークサービス利用者の行動を表す様々なデータが入手可能である。本稿では、このようなネットワーク利用者の行動に関するデータを「通信行動データ」と呼ぶ。通信行動データには、ネットワーク利用者の興味や社会のトレンドが反映されていると考えられている。そのため、通信行動データを利用して、ネットワーク利用者やそのコミュニティの状態を推定・予測する「コミュニティ計測」の研究が活発に行われている。通信行動データを利用したコミュニティ計測の代表的な例としては、感染症の流行の予測、グループの生産性の予測、グループにおける個人の重要度の推定、などが挙げられる。本稿では、様々な分野で行われているコミュニティ計測の研究動向を紹介し、さらに、その研究課題ならびに今後の展望について議論する。
キーワード　通信行動、コミュニティ計測、ソーシャルメディア、ソーシャルネットワーク

A Review on Community Instrumentation Techniques utilizing Communication Behavioral Data

Sho TSUGAWA†

† Faculty of Engineering, Information and Systems, University of Tsukuba
1–1–1 Tennodai, Tsukuba, Ibaraki 305–8573, Japan
E-mail: †s-tugawa@cs.tsukuba.ac.jp

Abstract　Large-scale and fine-grained records of network users' activities are currently available on the Internet. For instance, records of users' activities in several network services such as users' post on social media, review documents on e-commerce sites, and data of trends in web search engine, are available. In this paper, we refer such data as *communication behavioral data*. Communication behavioral data can reflect users' interest and trends in the society, and therefore, many researchers utilize the data for inferring or predicting the states of the network users and their communities. This paper focuses on such techniques called *community instrumentation*. Examples of community instrumentation include predicting disease spread, inferring group performance, and inferring importance of individuals in a group. Such techniques have been studied in wide-range of research areas. This paper introduces research trends of community instrumentation techniques, and also discusses their open issues and future directions.
Key words　Communication behavior, Community instrumentation, Social media, Social network

1 はじめに

様々なネットワークサービスの登場・普及により、ネットワーク利用者の行動に関する膨大かつ粒度の細かいデータが様々な場所に蓄積されている [1–4]。特に Web 上には一般に公開された膨大な量のデータが蓄積されている。例えば、代表的なソーシャルメディアである Twitter では、そのユーザの投稿とユーザ間の関係のデータが蓄積され、入手可能となっている。また、Amazon などの電子商取引サイトにおけるレビュー文書、Google トレンド (注1) のような検索エンジンにおけるトレンドデータ、Wikipedia における各ページの閲覧数など、ネットワークサービス利用者の行動を表す様々なデータが入手可能である。

(注1)：https://www.google.com/trends/

表 1　様々な通信行動データの例とその利用可能範囲および取得のコスト

階層	通信行動データの例	利用可能範囲	取得のコスト
利用者	脳波	計測対象の人の同意を得たもののみ	高 (専用の機器が必要)
	人の人脈	計測対象の人の同意を得たもののみ	高 (インタビューやアンケートが必要)
	人の満足度	計測対象の人の同意を得たもののみ	高 (インタビューやアンケートが必要)
サービス	Twitter の投稿	一般公開 (アクセスレートの制限あり)	低
	Wikipedia の記事閲覧数	一般公開	低
	Google における検索ログ	Google のみ	低
ネットワーク	サーバのログ	サーバの運用者のみ	低
	通信機器のログ	ネットワーク事業者のみ	低

また一般には公開されていないが、サーバやルータ、スイッチなどにも通信ログデータが蓄積されており、これらも、サービス事業者や通信事業者にとっては入手可能なネットワーク利用者の行動を表すデータである。

このようなネットワーク利用者の行動に関する「通信行動データ」が入手可能になってきたことから、これを様々な用途に応用する研究が活発に行われている [2,3,5,6]。例えば、通信行動データから利用者にとって適切な情報を推薦する情報推薦の研究は古くから行なわれており、この技術は、電子商取引サイトのユーザに対する商品推薦 [7]、やソーシャルメディアのユーザに対するフォローすべきユーザの推薦 [8] といったサービスに利用されている。

本稿では、通信行動データを利活用する研究の中でも特に、通信行動データから、ネットワーク利用者やそのコミュニティの状態を推定・予測する研究に着目する。通信行動データは、ネットワーク利用者の興味や社会のトレンドを反映していると考えられるため、通信行動データは人の関わる様々な現象や事象を推定・予測するのに有用であると期待されている [2–4,6,9]。これまで、通信行動データの一種であるソーシャルメディアにおける投稿が、感染症の流行の予測 [4] や、景気動向の予測 [10]、個人のうつ傾向の推定 [9,11] などに有用であることが示されている。

本稿では、人やそのコミュニティの状態を推定・予測する技術を「コミュニティ計測」と呼び、コミュニティ計測に関する研究の動向を概説することを目的とする。コミュニティ計測の研究は学際的であり、コンピュータサイエンスの分野の中でも、Web [2,11–15]、データマイニング [16,17]、人工知能 [18–20]、自然言語処理 [4,21,22]、ヒューマンコンピュータインタラクション [9,23–28]、など様々な分野で研究が行われている。また扱う対象が人や社会であるため、社会科学 [29] や経済学 [3] といったコンピュータサイエンス以外の分野でも研究が行われている。本稿では、このように様々な分野で推進されているコミュニティ計測に関する研究の整理を行う。さらに、コミュニティ計測の研究課題ならびに今後の展望についても議論する。

本稿の構成は以下の通りである。まず、2 章では、本稿で扱う通信行動データを紹介する。3 章では、通信行動データを利用したコミュニティ計測の研究を、計測の対象とするコミュニティの粒度で分類し、どのような目的で研究が行われているかを紹介する。4 章では、コミュニティ計測で用いられる主要な技術について解説する。5 章では、コミュニティ計測の研究課題および今後の展望について議論する。最後に、6 章において本稿のまとめを述べる。

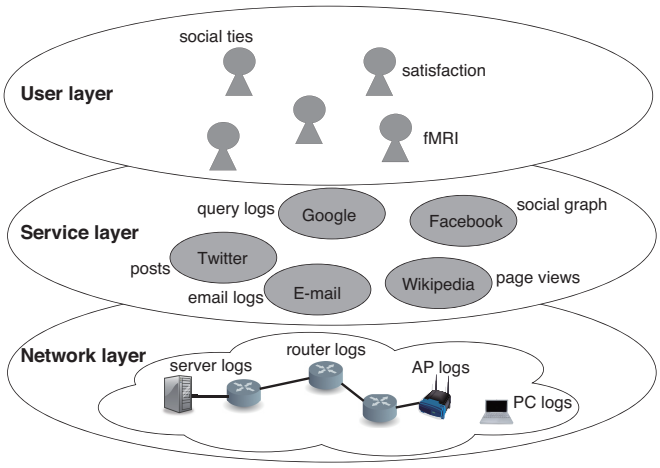

図 1: ネットワーク、サービス、およびそれらの利用者の 3 つの階層

2　様々な通信行動データ

ネットワーク利用者の行動に関する通信行動データには、様々な種類が存在する。図 1 は、ネットワークとその上で提供されるサービス、およびそれらの利用者の 3 つの階層を表している。図 1 に示す、各階層ごとに様々な通信行動データが存在する。表 1 に、ネットワーク、サービス、および利用者の各階層とその階層における通信行動データの例を示す。このように、通信機器から取得するログデータから、利用者から取得する脳波のデータまで様々な種類の通信行動データが存在する。

これらの通信行動データの利用可能な範囲や、データを取得すること自体のコストは、データによって異なる。表 1 には各通信行動データの利用可能範囲およびデータを取得することのコストも示している。利用者に関する通信行動データの取得には、その人へのインタビューや専用の計測機器が必要であるため、そのようなデータを取得すること自体が高コストである。また、その人の同意を得られた場合のみしか取得したデータを利用することができないため、利用範囲も制限される。一方で、ネットワークにおける各種の通信ログやサービスの利用履歴は、ネットワークやサービスを運用する上で元々蓄積しているものであるため、データを取得すること自体のコストは比較的低い。ただし、ネットワークにおける通信行動データは、利用者のプライバシーの観点から公開できないものも多く、例えばインターネットサービスプロバイダ (ISP) の有する通信記録は一般に ISP 以外が知ることはできない。それに対して、ソーシャルメディアにおける投稿や、Wikipedia の各ページへのアク

表 2　コミュニティ計測の研究の計測対象と目的、および用いるデータ基づく分類

		用いるデータ		
	推定/予測の対象	ソーシャルメディア	Wikipedia の記事閲覧数	Web 検索履歴
マクロレベル	社会における病気の流行	[4, 33]	[34]	[35]
	経済動向	[10, 12, 13, 16, 21, 22, 29]		[3]
	選挙結果	[36, 37]		[38]
	旅行者数	[39]		[3]
メゾレベル	地域ごとの健康に関する統計	[14]		
	地域ごとの幸福度	[23, 40]		
	商品/サービスの売り上げ	[2, 17, 18]	[19, 41]	[3, 19, 42]
	グループの生産性や成功度	[24, 25, 43, 44]		
ミクロレベル	個人の健康状態	[9, 11, 20, 26, 27, 45]		
	個人のグループにおける貢献度/重要度	[15, 28]		

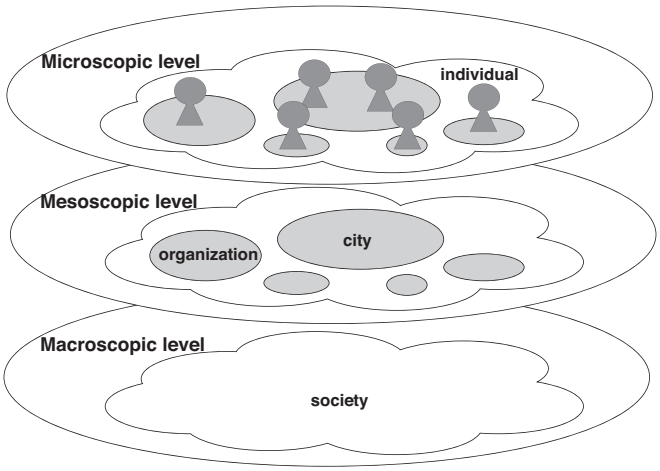

図 2: 人の社会の階層構造

セス数など、Web サービスのデータの中には一般に公開され、利用可能なデータも存在している。

本稿では特に、一般に公開されており入手の容易な通信行動データである Web サービスの利用履歴を利用し、人やコミュニティの状態を推定するコミュニティ計測の研究に着目する。なお、ネットワーク事業者やサービス提供者のみが有する通信行動データを利用したコミュニティ計測については本稿の対象外であるが、このような研究については、例えば、携帯電話から得られるデータを利用したコミュニティ計測技術の研究動向を解説した文献 [30] などを参照されたい。また、利用者から直接得られる通信行動データを利用した研究の例としては、文献 [31, 32] などが挙げられる。

3　コミュニティ計測の対象と目的

3.1　コミュニティ計測の研究の分類

コミュニティ計測の対象となる人や人のコミュニティは階層的な構造を有している (図 2)。図 2 に示すように人の社会は 1 つの大きなコミュニティであるとみなすことが可能であり、人の社会は組織や特定の都市などさらに小さなコミュニティで構成されている。コミュニティを構成する最小の単位は人である。このように観測するスケールによって、我々の社会には様々な粒度のコミュニティが存在する。

本稿では、コミュニティ計測の研究を、この階層と対応させ、国や社会といった「マクロレベル」のコミュニティを対象とした研究、組織や特定の地域などの「メゾレベル」のコミュニティを対象とした研究、そしてコミュニティを構成する個人を対象とした「ミクロレレベル」の研究、に分類する。既存のマクロレベル、メゾレベル、ミクロレベルのコミュニティ計測の研究をそれぞれの目的及び用いるデータに基づいて分類した結果を表 2 に示す。以降では、このマクロレベル、メゾレベル、ミクロレベルの分類に従って、コミュニティ計測の代表的な研究例を紹介する。なお、紙面の都合上以下で紹介できない文献についても、表 2 には掲載している。

3.2　マクロレベルのコミュニティ計測

マクロレベルのコミュニティ計測の研究は、人の社会や国などの大きな単位のコミュニティの状態を推定・予測することを目的としている。

マクロレベルのコミュニティ計測の研究における主要なタスクの一つとして、インフルエンザなどの感染症の流行状況の推定・予測 [4, 33–35] が挙げられる。従来から、感染症の流行状況は、日本の国立感染症研究所や、米国における疾病予防管理センター (CDC) によって、病院の診療情報を元に推定、公表されてきた。このような調査方法は、データの収集にコストがかかり、推定結果の公表までに時間を要することから、通信行動データを用いて、このような調査を代替する手法の研究が行われている [4, 35]。Ginsberg らは、Google における検索ログから、インフルエンザの流行状況を推定する手法を提案している [35]。これは、Google Flu Trends (GFT) (注2) として、実際にサービスとしても運用されている。Aramaki らは、Google における検索ログのような非公開データからではなく、Twitter におけるインフルエンザに関する投稿からインフルエンザの流行状況を推定する手法を提案している [4]。さらに、現在の流行状況だけでなく、将来の流行状況を予測する手法の研究も始まっている [34]。ただし、GFT による推定と、CDC の実際の発表とが 2 倍程度ずれる例が報告されるなど、通信行動データからの予測で従来の調査を置き換えることの問題も指摘されている [46]。

感染症の流行状況以外にも、将来の株価 [10, 13, 16, 21, 22, 29] や選挙の結果 [36–38] など、社会における政治や経済の動向も主要な予測の対象である。このような研究では、Twitter などのソーシャルメディアにおける投稿を活用する方法が注目されている。Zhang らは Twitter における投稿を用いて、Dow Jones

(注2)：https://www.google.org/flutrends/about/

や NASDAQ などの株価指標を推定している [29]。Nguyen らは株の個別銘柄に関する掲示板の投稿から個別の株価の上下を予測する手法を提案している [22]。また、Twitter における投稿から、選挙結果を予測した様々な事例も報告されている [36,37]。

3.3 メゾレベルのコミュニティ計測

メゾレベルのコミュニティ計測の研究では、特定の地域やグループなど社会や国と比べて小さな単位のコミュニティの状態を推定・予測することを目的としている。

メゾレベルのコミュニティ計測における主要なタスクとして、商品やサービスの売上予測が挙げられる。Goel らは、映画やビデオゲーム、音楽などの売上を検索ログから推定する手法を提案している [42]。Asur らは、ソーシャルメディアにおける投稿から映画の売上を予測している [2]。角田らは Google Trends と Wikipedia の記事の閲覧数を用いて、自動車の販売台数を予測する手法を提案している [19]。

一方、組織やグループの生産性を推定する研究も行われている。Munson らは学生のプロジェクト内でのメッセージ交流の履歴からそのプロジェクトの生産性を推定するのに有効な特徴量を調査している [24]。Tsay らは GitHub におけるソフトウェア開発者の活動履歴から、開発プロジェクトの成功度を推定する手法を提案している [25]。

メゾレベルのコミュニティ計測の研究では、計測対象に関する通信行動データを、マクロレベルの計測の研究と比較して、少量しか得ることができず、それが問題となる場合も存在する。文献 [40] では位置情報付きツイートを用いて、アメリカの各州に関する統計を予測しているが、例えば日本の県程度の狭い単位で同様の予測を行う場合には、予測に用いる通信行動データの量が不足するという問題が生じ得る。また、ソフトウェア開発の成功度を推定する研究において、ある程度の量のメッセージ交流を行うコミュニティでないと、その成功度を推定できないことが報告されている [44]。

3.4 ミクロレベルのコミュニティ計測

ミクロレベルのコミュニティ計測の研究では、コミュニティを構成する個人の状態を推定・予測することを目的としている。

ミクロレベルのコミュニティ計測の研究の中でも、個人のグループにおける重要度や貢献度を推定する研究は比較的多く行われている。例えば、Zhang らは質問応答コミュニティにおけるユーザ間のメッセージ交流の履歴から、ユーザの重要度を推定する手法を提案している [15]。また、近年の研究としては、企業内の SNS の履歴から社員の貢献度を推定する研究も行なわれている [28]。

個人の健康状態の推定も注目を集めている研究領域である。Sadilek では、ソーシャルメディアにおける活動履歴から、ユーザのインフルエンザへの感染の有無を推定する手法を提案している [20]。Choudhury ら [11] および Tsugawa ら [9] は、ソーシャルメディアにおける活動履歴からユーザのうつ傾向を推定する手法を提案している。

ミクロレベルのコミュニティ計測の研究では、メゾレベルのコミュニティ計測の研究と比べてもさらに、計測対象に関する通信行動データを得ることが難しいことが問題となっている。Shami ら [28] および、Tsugawa ら [9] は、ミクロレレベルのコミュニティ計測において必要とされる通信行動データの量についても議論しており、ある程度のデータ量が得られない場合には、計測の精度が著しく低下することを報告している。

4 コミュニティ計測に用いられる技術

4.1 コミュニティ計測の流れ

コミュニティ計測は、基本的に、コミュニティの状態に関する何らかの変数を予測、あるいは推定する教師あり機械学習の問題として定式化される。まず通信行動データからコミュニティ計測に有用な特徴量を抽出する。抽出した特徴量と教師データを用いた機械学習によってコミュニティの状態を推定・予測するモデルを構築する。以降では通信行動データから予測に有用な特徴を抽出する技術、及び抽出した特徴量から予測モデルを構築するための技術について説明する。

4.2 通信行動データからの特徴抽出

通信行動データからの特徴抽出の方法は目的に依存して様々存在するが、特に自然言語処理やネットワーク分析の技術が広く用いられている。

ソーシャルメディアにおける投稿を用いる研究においては、多くの場合、自然言語処理の技術を用いて、大量の投稿から有用な特徴の抽出を行う。ソーシャルメディアにおける投稿を解析し、特定の単語の出現頻度や特定の単語の含まれる投稿数などの特徴量を抽出することは多くの研究で行われている [11,12,29]。ただし、この方法で、大量のデータから予測に有用な単語を選ぶためには、計測対象に関するドメイン知識が必要となる。それに対して、トピックモデル [47] と呼ばれる技術による特徴抽出も注目を集めている。トピックモデルは、文書 (例えばブログ記事) には、いくつかのトピックが存在し、そのトピックに基づいてその文書中の単語が生成されているという考えに基づき、文書のトピック比率を推定する [47]。単語の出現頻度よりも、トピック比率の方が、コミュニティ計測のための特徴量として有用であるという結果も報告されている [9,22]。なお、トピックモデルについての解説は、文献 [48] などが詳しい。また、文書に書かれている内容がポジティブなものであるか、ネガティブなものであるかという評価極性を推定する評判分析の技術も、コミュニティ計測の研究において広く利用されている [12, 14, 21–23]。評判分析では、辞書や文書中での単語の共起関係などを元に、単語、文、文書など様々な単位でその評価や感情の極性を推定する。評判分析についての解説は、文献 [49] などを参照されたい。

様々なモノとモノとの関係を表現したネットワークから特徴を抽出するネットワーク分析の技術も有用な特徴抽出の手法である。Zhang らは、コミュニティにおける参加者間の交流関係からソーシャルネットワークを構築し、PageRank や HITS などのアルゴリズムを用いて、参加者の重要度を推定している [15]。また、Tsugawa ら [44] は、開発プロジェクト参加者間のソーシャルネットワークから求めた中心性指標から、プロジェクト参加者のリーダシップの強さを推定している。さらに推定したリーダシップの強さが、プロジェクトの成功度の推定に有用であることを示している [44]。また、Corley らはインフルエンザの流行予測に有用なブログサイトの特定に、ブログネットワークを用いている [33]。このようなネットワーク分析手法の解説については、文献 [50] などを参照されたい。

4.3 機械学習による予測モデルの構築

コミュニティ計測の研究では、通信行動データから抽出され

た特徴量を入力とし、様々な機械学習の手法を用いて予測モデルを構築する。ここでは、回帰分析のような基本的な手法 [2,28] から、SVM (Support Vector Machine) などの高次元データの分類に適した学習手法も用いられる [4, 9, 11, 20]。また、株価や病気の流行度など、時系列データを予測する研究においては、時系列解析の手法が用いられる [3, 19]。ここでも、予測対象のドメイン知識を活用する場合も存在し、例えば、感染症のモデルである SIR モデルを拡張することにより、インフルエンザの流行度を予測する手法が提案されている [34]。なお、 Deep Learning と呼ばれる多層のニューラルネットワークを用いた機械学習手法を用いて、個人のストレス度を推定する手法も提案されている [45]。Deep Learning を用いたコミュニティ計測の研究例はまだそれほど多くないが、これは、画像認識や音声認識の分野で注目を集めている技術であり、コミュニティ計測においても有効な手法ではないかと注目している。

5 課題と今後の展望

5.1 課題：コミュニティ計測の利用方法の確立

コミュニティ計測の研究は活発に行われているものの、これを実際どのように利用していくのかについては、未だ議論の途中段階である。Choudhury らは、ソーシャルメディアからうつ傾向を推定することが技術的に可能になったとして、推定結果を提示することの倫理的な問題への議論が必要であると述べている [27]。Lazer らは、通信行動データを用いた予測技術は、他の従来の計測技術を置き換えるのではなく、相補的に用いるべきであると述べている [46]。我々も、コミュニティ計測の技術は、その計測対象を理解する上での補助として、追加の情報を提供するものであると考えている。そのため、計測の精度の向上や適用範囲の拡大に加えて、どのように利用していくのかの検討が重要な課題であると考えている。

5.2 課題：データ量とノイズのトレードオフ

3 章でも議論したように、マクロレベル、メゾレベル、ミクロレベルと、より粒度の細かい対象の状態を計測しようとすると、利用可能な通信行動データの量がより少なくなってしまう。一方、粒度の粗い対象を計測する場合には、大量のデータが手に入るものの、非常にノイズの多いデータから有用な特徴量を抽出することが難しいという課題が報告されている [37]。大量データに存在するノイズの問題については、ノイズ除去のための方法について論文中に詳しく記載するなど、ノウハウの共有も行なわれている [37]。一方、少量データが問題となることは、特にミクロレベルのコミュニティ計測の研究において指摘されているものの [9, 28]、有効な解決方法は我々の知る限り未だ提案されていない。

5.3 今後の展望：アクティブ計測に相当する技術

通信ネットワークの計測における「アクティブ計測」に対応するコミュニティ計測技術の研究は、有望な研究の方向の 1 つであると考えている。ネットワーク計測では、ログを利用するパッシブ計測と、実際にパケットを流すアクティブ計測が存在する。本稿で紹介した研究は、全てパッシブ計測に分類される技術の研究である。計測対象に関するデータが非常に少ない状況においては、アクティブ計測のようなアプローチも有効ではないかと期待している。たとえば、ユーザの心理状態を推定したい場合に、ユーザ端末に何らかの情報提示を行い、それに対する行動の変化を計測するなどの方法が考えられる。

5.4 今後の展望：クロスドメイン/クロスレイヤの連携

ドメインおよび階層をまたがる通信行動データの相補的な利用も、有望な研究の方向であると考えている。本稿で紹介した研究では、単一の種類のデータを利用することが多く、複数のデータを組み合わせて利用する例はまだ少ない。異なる種類のデータ、さらには図 1 における異なる階層の通信行動データを利用することで、特にミクロレベルの計測で問題となっている少量データの問題が解決できるのではないかと考えている。

なお、マサチューセッツ工科大学の Reality Minig Project のように、人の装着したセンサ情報や電子メールの履歴など複数の通信行動データを利用して、組織の状態を推定する研究も既に行われている [31, 32]。例えば、文献 [31] では、バッジ型のセンサから取得した対面での交流の履歴と、電子メールの履歴を用いて、組織におけるグループの生産性や個人の満足度などを推定している。ただし、このような研究では、同じデータを複数の研究者が用いて、より良い手法を開発するということが行いにくいため、個別の研究事例の報告がほとんどである。

今後、異種の通信行動データを用いたコミュニティ計測の研究を発展させていくためには、研究者間で共有されたデータを用いた議論が重要になると考えている。このようなデータの共有の例は既に存在し、欧州における SocioPatterns プロジェクト (注3) では、センサを用いて計測した個人の対面交流の履歴を匿名化して公開している。Data for Development (D4D) (注4) では、途上国における携帯電話の利用履歴を公開している。このような通信行動データを共有する取り組みによって、クロスドメイン/クロスレイヤの通信行動データを組み合わせたコミュニティ計測の研究が加速するものと期待している。

6 ま と め

本稿では、ネットワーク利用者の行動に関する通信行動データを用いて、人や人のコミュニティの状態を推定・予測するコミュニティ計測の技術に関する研究動向を紹介した。さらに、通信行動データを用いたコミュニティ計測の研究課題と今後の展望についても議論した。

謝辞 本研究の一部は JSPS 科研費 26870076、ならびに電気通信普及財団の支援を受けている。

文 献

[1] D. J. Watts, “A twenty-first century science,” *Nature*, vol. 445, no. 7127, p. 489, 2007.

[2] S. Asur, B. Huberman *et al.*, “Predicting the future with social media,” in *Proc. of WI-IAT*, 2010, pp. 492–499.

[3] H. Choi and H. Varian, “Predicting the present with Google trends,” *Economic Record*, vol. 88, no. s1, pp. 2–9, 2012.

[4] E. Aramaki, S. Maskawa, and M. Morita, “Twitter catches the flu: detecting influenza epidemics using Twitter,” in *Proc. of EMNLP*, 2011, pp. 1568–1576.

[5] W. Fan and M. D. Gordon, “The power of social media analytics,” *Communications of the ACM*, vol. 57, no. 6, pp. 74–81, 2014.

[6] D. Gayo-Avello, “A meta-analysis of state-of-the-art electoral prediction from Twitter data,” *Social Science Computer Review*, vol. 31, no. 6, pp. 649–679, 2013.

[7] G. Linden, B. Smith, and J. York, “Amazon.com recommenda-

(注3)：http://www.sociopatterns.org/about/

(注4)：http://www.d4d.orange.com/en/Accueil

tions: Item-to-item collaborative filtering," *IEEE Internet Computing*, vol. 7, no. 1, pp. 76–80, 2003.

[8] P. Gupta, A. Goel, J. Lin, A. Sharma, D. Wang, and R. Zadeh, "WTF: The who to follow service at Twitter," in *Proc. of WWW'13*, 2013, pp. 505–514.

[9] S. Tsugawa, Y. Kikuchi, F. Kishino, K. Nakajima, Y. Itoh, and H. Ohsaki, "Recognizing depression from Twitter activity," in *Proc. of CHI*, 2015, pp. 3187–3196.

[10] J. Bollen, H. Mao, and X. Zeng, "Twitter mood predicts the stock market," *Journal of Computational Science*, vol. 2, no. 1, pp. 1–8, 2011.

[11] M. De Choudhury, M. Gamon, S. Counts, and E. Horvitz, "Predicting depression via social media," in *Proc of ICWSM*, 2013, pp. 128–137.

[12] B. O'Connor, R. Balasubramanyan, B. R. Routledge, and N. A. Smith, "From tweets to polls: Linking text sentiment to public opinion time series," in *Proc. of ICWSM*, 2010, pp. 122–129.

[13] M. De Choudhury, H. Sundaram, A. John, and D. D. Seligmann, "Can blog communication dynamics be correlated with stock market activity?" in *Proc. of HT*, 2008, pp. 55–60.

[14] M. De Choudhury, S. Counts, and E. Horvitz, "Social media as a measurement tool of depression in populations," in *Proc. of WebSci*, 2013, pp. 47–56.

[15] J. Zhang, M. S. Ackerman, and L. Adamic, "Expertise networks in online communities: structure and algorithms," in *Proc. of WWW*, 2007, pp. 221–230.

[16] E. J. Ruiz, V. Hristidis, C. Castillo, A. Gionis, and A. Jaimes, "Correlating financial time series with micro-blogging activity," in *Proc. of WSDM*, 2012, pp. 513–522.

[17] D. Gruhl, R. Guha, R. Kumar, J. Novak, and A. Tomkins, "The predictive power of online chatter," in *Proc. of KDD*, 2005, pp. 78–87.

[18] G. Mishne and N. S. Glance, "Predicting movie sales from blogger sentiment," in *Proc. of the AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, 2006, pp. 155–158.

[19] 角田孝昭, 吉田光男, 津川翔, 山本幹雄, "状態空間モデルを用いた検索トレンドとページビューからの自動車販売台数の予測," 人工知能学会全国大会論文集, vol. 29, pp. 1–4, 2015.

[20] A. Sadilek, H. A. Kautz, and V. Silenzio, "Predicting disease transmission from geo-tagged micro-blog data." in *Proc. of AAAI*, 2012, pp. 136–142.

[21] J. Si, A. Mukherjee, B. Liu, Q. Li, H. Li, and X. Deng, "Exploiting topic based Twitter sentiment for stock prediction," in *Proc. of ACL*, 2013, pp. 24–29.

[22] T. H. Nguyen and K. Shirai, "Topic modeling based sentiment analysis on social media for stock market prediction," in *Proc. of ACL*, 2015, pp. 1354–1364.

[23] D. Quercia, J. Ellis, L. Capra, and J. Crowcroft, "Tracking gross community happiness from tweets," in *Proc. of CSCW*, 2012, pp. 965–968.

[24] S. A. Munson, K. Kervin, and L. P. Robert Jr, "Monitoring email to indicate project team performance and mutual attraction," in *Proc. of CSCW*, 2014, pp. 542–549.

[25] J. T. Tsay, L. Dabbish, and J. Herbsleb, "Social media and success in open source projects," in *Proc. of CSCW Companion*, 2012, pp. 223–226.

[26] M. De Choudhury, S. Counts, E. J. Horvitz, and A. Hoff, "Characterizing and predicting postpartum depression from shared Facebook data," in *Proc. of CSCW*, 2014, pp. 626–638.

[27] S. Chancellor, Z. J. Lin, E. L. Goodman, S. Zerwas, and M. De Choudhury, "Quantifying and predicting mental illness severity in online pro-eating disorder communities," in *Proc. of CSCW to appear*, 2016.

[28] N. S. Shami, M. Muller, A. Pal, M. Masli, and W. Geyer, "Inferring employee engagement from social media," in *Proc. of CHI*, 2015, pp. 3999–4008.

[29] X. Zhang, H. Fuehres, and P. Gloor, "Predicting stock market indicators through Twitter "I hope it is not as bad as I fear"," *Procedia-Social and Behavioral Sciences*, vol. 26, pp. 55–62, 2011.

[30] W. Z. Khan, Y. Xiang, M. Y. Aalsalem, and Q. Arshad, "Mobile phone sensing systems: A survey," *IEEE Communications Surveys & Tutorials*, vol. 15, no. 1, pp. 402–427, 2013.

[31] D. O. Olguín, B. N. Waber, T. Kim, A. Mohan, K. Ara, and A. Pentland, "Sensible organizations: Technology and methodology for automatically measuring organizational behavior," *IEEE Trans. on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 39, no. 1, pp. 43–55, 2009.

[32] N. Eagle and A. Pentland, "Reality mining: Sensing complex social systems," *Personal and Ubiquitous Computing*, vol. 10, no. 4, pp. 255–268, 2006.

[33] C. D. Corley, D. J. Cook, A. R. Mikler, and K. P. Singh, "Text and structural data mining of influenza mentions in web and social media," *International Journal of Environmental Research and Public Health*, vol. 7, no. 2, pp. 596–615, 2010.

[34] K. S. Hickmann, G. Fairchild, R. Priedhorsky, N. Generous, J. M. Hyman, A. Deshpande, and S. Y. Del Valle, "Forecasting the 2013–2014 influenza season using Wikipedia," *PLoS Computational Biology*, vol. 11, no. 5, p. e1004239, 2015.

[35] J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant, "Detecting influenza epidemics using search engine query data," *Nature*, vol. 457, no. 7232, pp. 1012–1014, 2009.

[36] A. Tumasjan, T. Sprenger, P. Sandner, and I. Welpe, "Predicting elections with Twitter: What 140 characters reveal about political sentiment," in *Proc of ICWSM*, 2010, pp. 178–185.

[37] A. Khatua, A. Khatua, K. Ghosh, and N. Chaki, "Can #Twitter_Trends predict election results? evidence from 2014 Indian general election," in *Proc. of HICSS*, 2015, pp. 1676–1685.

[38] L. Granka, "Using online search traffic to predict us presidential elections," *PS: Political Science & Politics*, vol. 46, no. 02, pp. 271–279, 2013.

[39] D. Barchiesi, H. S. Moat, C. Alis, S. Bishop, and T. Preis, "Quantifying international travel flows using Flickr," *PloS ONE*, vol. 10, no. 7, p. e0128470, 2015.

[40] J. Loff, M. Reis, and B. Martins, "Predicting well-being with geo-referenced data collected from social media platforms," in *Proc. of SAC*, 2015, pp. 1167–1173.

[41] M. Mestyán, T. Yasseri, and J. Kertész, "Early prediction of movie box office success based on Wikipedia activity big data," *PloS ONE*, vol. 8, no. 8, p. e71226, 2013.

[42] S. Goel, J. M. Hofman, S. Lahaie, D. M. Pennock, and D. J. Watts, "Predicting consumer behavior with Web search," *Proceedings of the National Academy of Sciences*, vol. 107, no. 41, pp. 17 486–17 490, 2010.

[43] X. Yang, D. Hu, and D. M. Robert, "How microblogging networks affect project success of open source software development," in *Proc. of HICSS*, 2013, pp. 3178–3186.

[44] S. Tsugawa, H. Ohsaki, and M. Imase, "Inferring leadership of online development community using topological structure of its social network," *Journal of the Infosocionomics Society*, vol. 7, no. 1, pp. 17–27, 2012.

[45] H. Lin, J. Jia, Q. Guo, Y. Xue, Q. Li, J. Huang, L. Cai, and L. Feng, "User-level psychological stress detection from social media using deep neural network," in *Proc. of the ACM International Conference on Multimedia*, 2014, pp. 507–516.

[46] D. Lazer, R. Kennedy, G. King, and A. Vespignani, "The parable of Google Flu: traps in big data analysis," *Science*, vol. 343, pp. 1203–1205, 2014.

[47] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *The Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[48] D. M. Blei, "Probabilistic topic models," *Communications of the ACM*, vol. 55, no. 4, pp. 77–84, 2012.

[49] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Foundations and Trends in Information Retrieval*, vol. 2, no. 1-2, pp. 1–135, 2008.

[50] S. Wasserman and K. Faust, *Social network analysis: Methods and applications*. Cambridge university press, 1994, vol. 8.