情報処理学会論文誌 Vol.56 No.10 2044–2053 (Oct. 2015)

Twitterにおける候補者の選挙地盤に着目した国政選挙の当選者予測

那須野 薫[1,a)] 奥山 晶二郎[2] 中西 鏡子[2] 松尾 豊[1,b)]

受付日 2014年8月16日, 採録日 2015年7月1日

概要：近年，Twitter のデータを用いて選挙結果の予測を試みる研究の報告が活発である．本研究では，選挙結果を高い精度で予測するモデルの構築を目指し，社会学で古くから選挙当落の重要な要素の 1 つとされてきた選挙地盤を定量的に測定，指標化し，この指標を用いることで既存手法の拡張を試みる．選挙地盤に関する指標は選挙地盤のリーチ，バラエティ，ロイヤルティという 3 つの指標を提案する．選挙運動へのインターネットの利用が初めて解禁された 2013 年の参議院議員選挙を対象とした評価実験の結果，本研究で提案した 3 つの選挙地盤に関する指標は選挙結果の予測に有効であることが示された．また，本研究で用いた手法は既存手法と比較して F 値が約 70%高く，選挙運動への Twitter の活用は選挙結果に小さいものの影響があることが示唆された．

キーワード：当選者予測，Twitter，選挙地盤，国政選挙

Predicting Japanese General Election by Focusing on Candidates' Constituency on Twitter

Kaoru Nasuno[1,a)] Shojiro Okuyama[2] Kyoko Nakanishi[2] Yutaka Matsuo[1,b)]

Received: August 16, 2014, Accepted: July 1, 2015

Abstract: Studies for predicting or examining election results using Twitter data becomes popular recently. In this paper, aiming at making a prediction model with high accuracy, we extend a previous method by using features representing candidates' constituency which has been considered as one of the most significant elements on election result in sociology. We propose three indicators as features on constituency: reach of constituency, variety of constituency and loyalty of constituency. Evaluation test is conducted with Twitter data during the period of Japanese general election in 2013. The result of evaluation test shows the three indicators we propose are useful for electoral prediction. Besides F-measure by our method is shown to be higher by about 70% than that by the previous method and this indicates that using Twitter for electoral campaign might have an effect on election results.

Keywords: election prediction, Twitter, candidates' constituency, Japanese general election

1. はじめに

選挙運動では，選挙当落に重要な要素である地盤，看板，鞄のいわゆる「三バン」が重視されてきた [1]．すなわち，選挙の当選には候補者の選挙区内の後援や支持が強固である（地盤が固い）こと，候補者の知名度が高い（市中の看板のように知られている）こと，候補者の選挙資金が豊富である（鞄が札束で一杯である）ことが重要であると考えられてきた．

近年，ますます多くの有権者がソーシャルメディアを利用するようになったため，候補者も知名度の向上や選挙地盤の強化などを期待して，選挙運動にソーシャルメディアを利用するようになってきた．国内で初めてインターネッ

[1] 東京大学
The University of Tokyo, Bunkyo, Tokyo 113–8656, Japan
[2] 朝日新聞社デジタル編集部
Interactive Media & News Section, The Asahi Shimbun, Chuo, Tokyo 104–8011, Japan
a) nasuno@weblab.t.u-tokyo.ac.jp
b) matsuo@weblab.t.u-tokyo.ac.jp

トの利用が解禁された 2013 年の参議院議員選挙においても少なくない候補者が選挙期間中にソーシャルメディアを利用し有権者に働きかけていた．

ソーシャルメディアにおけるユーザの活動は現実の活動とは異なりデータが取得しやすいということもあり，ソーシャルメディアにおけるユーザの活動データを利用して社会動向の説明や予測を試みる研究が活発に行われている．マイクロブログサービスの Twitter はこうした目的に広く用いられており，140 文字以下の投稿であるツイートを解析することで，インフルエンザの流行予測 [2] や株式市場の動向予測 [3] が可能であることが示されている．また，Twitter からのデータを利用して選挙結果の予測を試みる研究も少なくない [4], [5], [6], [7], [8]．

特に，Cameron らの研究 [8] では，フォロワの数が多ければ候補者がよく認知されており選挙に当選する確率が高いという考えに基づき，候補者のフォロワの数やフォロワの数の推移から選挙当落の予測が試みられたが，予測実験では高い精度は得られていなかった．結果が芳しくなかった原因として，いくつか考えられるが，たとえば，フォロワの数だけでは候補者の知名度をうまく表現できないのかもしれない．もしくは，知名度だけでなく選挙地盤の強さや選挙資金の豊富さを考慮することが重要なのかもしれない．いずれにせよ，候補者のフォロワの数だけを考慮したモデルでは選挙当落の予測は難しいと考えられる．

今後，選挙運動へのソーシャルメディア活用の活発化が期待されるなかで，当選しやすいユーザの特徴や当選に寄与するユーザの行動に関する知見を獲得したり，精度の高い予測モデルを構築したりすることは候補者や候補者を選定する政党にとって有用であると考えられる．なぜなら，候補者にとっては，当選に寄与する活動を中心に取り組むことで効率的に選挙運動を行うことができる可能性があり，また，政党にとっては，候補者を選定する際にすでにソーシャルメディア上で当選しやすいと予測される人のなかから選定することで自党の議席を効率的に増やせる可能性があるからである．

本研究は，国政選挙の当選者を高い精度で予測するモデルを構築することを目指し，予測精度の向上に寄与するユーザの特徴や行動に関する知見を獲得することを目的とする．Cameron らの手法 [8] を拡張し，知名度の表現にフォロワの数だけでなく知名度を表現する可能性のある他の指標（フレンド数やアカウント認証の有無など）もあわせて利用することで予測精度の向上を試みる．また，それに加えて候補者の選挙地盤を表現するような指標（リーチ，バラエティ，ロイヤルティ）を提案し予測モデルに利用することで，さらに予測精度が向上するかを検証する．

評価実験では初めてインターネットの利用が解禁された 2013 年の参議院議員選挙に出馬した候補者の Twitter アカウントを対象に，教師あり学習により当選者の予測を試みる．予測モデルの構築には候補者のプロフィールに関するデータや選挙期間中の候補者のツイート 42,645 とそのリツイート 368,694 を用い，教師あり学習のアルゴリズムには学習後に各素性の重みを確認でき，また広く用いられ良好な結果が得られている Random Forest を用いる．

本研究の主な貢献は下記のとおりである．

- 社会学でこれまで議論されてきた選挙地盤を，ソーシャルメディア上で定量的に測定した事例である．
- Twitter における候補者の選挙地盤に着目した指標が選挙結果の予測に有効であることを示した．
- 選挙運動への Twitter の活用は選挙結果に小さいものの影響があることを示唆した．
- 選挙地盤が強固であるということは，Twitter において候補者やその後援者が排他的にツイートを広められるユーザの数が多い状態に相当することを示唆した．
- 政党名や候補者名に言及したツイートの分析などとあわせることで，予測精度が向上する可能性があることを示唆した．

本稿の構成は下記のとおりである．まず，次章で本研究の位置づけを明確にするために関連研究を説明する．3 章で Cameron らの手法の拡張方法について述べ，4 章で本研究の評価実験で用いるデータの取得方法や概観について述べ，5 章で評価実験を通して Cameron らの手法に加えた 2 つの拡張により予測精度が改善されるかを検証する．6 章で本研究の限界を整理し課題と今後の拡張可能性を述べ，7 章でまとめる．

2. 関連研究

本章では，本研究と既存研究の位置づけを明確にするために関連研究を整理する．Twitter のデータを用いた社会動向の予測研究における選挙結果の予測研究の位置づけや特徴を整理し，本研究の拡張対象である Cameron らの研究について説明する．

Twitter は日本でも多くの人が利用しているマイクロブログサービスで，Twitter のデータは社会動向の分析のために広く用いられている [9], [10]．Twitter のデータを利用して社会動向の予測に成功したという研究 [2], [3] がいくつか報告され他の分野への応用が研究される中で，特に，選挙という分野では一般性や精度の高い予測モデルの構築ができていないのが現状である．有権者に焦点を当てる研究 [4], [6], [7], [11], [12] と候補者に焦点を当てる研究 [8] に分けてそれぞれの特徴を整理すると，有権者に焦点を当てる研究は多く報告されているが，政党名に言及する有権者のツイート数に着目した Tumasjan らの研究 [4] を否定する研究 [6] を Jungherr らが報告し，候補者名や選挙関連語などを含む有権者のツイートの感情に着目した O'Connor らの研究 [7] で提案された手法や同様の手法を否定したり有効性を限定したりする研究を Chung ら [12] や Gayo-Avello

ら [11] が報告しており，一般性や精度の高い予測モデルの構築ができていないのが現状である．一方で，候補者に焦点を当てる研究はあまり多く報告されておらず，また，報告されている Cameron らの研究 [8] においても精度の高い予測モデルは得られていない．

本研究はこのような背景の中で Cameron らの研究を拡張することで，精度の高い予測モデルの構築を目指すものである．Cameron らの研究は，フォロワ数が多ければ候補者がよく認知されており選挙に当選する確率が高いという考えに基づいていると解釈でき，候補者のフォロワ数やその推移から予測モデルを構築し選挙当落の予測が試みられたが，実験では高い精度は得られていなかった．候補者のフォロワの数は選挙日から 2 カ月前，1 カ月前，1 週間前，当日の 4 時点のデータが用いられ，また予測モデルの構築にはロジスティック回帰が利用されていた．実験では，フォロワの数だけを考慮しても予測精度はほとんど得られておらず，Twitter のフォロワの数やその推移はあまり当選可能性と関係がないと結論づけた．また，先行研究 [13] をふまえてソーシャルネットワークの構造を考慮することで予測精度が向上する可能性を示唆した．

本章では，本研究の位置づけを明確にするため関連研究と拡張対象の研究について説明した．次章では，選挙結果の予測精度を向上させるための拡張法を説明する．

3. 拡張法

本章では，選挙結果の予測精度の向上を目指して本研究で実施する Cameron らの研究の拡張について説明する．本研究の手法と Cameron らの手法の共通要素は候補者に関する指標を素性とし教師あり学習のアルゴリズムによりモデルを構築し選挙当落の二値分類を行うという点であり，独自要素は予測精度を向上させるためにモデルの構築に用いる指標を拡張しているという点である．まず，指標の拡張を 2 段階で行うことを説明し，その後，それぞれの拡張法を詳細に述べる．

Cameron らの手法は候補者のフォロワの数を素性として選挙当落を予測するモデルを構築するというものであった．フォロワの数は候補者を認知し発言を受け取るユーザの数であることから，厳密に区別することは難しいものの，三バンの中で特に候補者の看板（知名度）に着目した手法であったと考えられる．これに加えて，三バンが選挙当落において重要な要素であることと候補者の鞄（選挙資金の豊富さ）は Twitter のデータからの推定が難しいと考えられることをふまえると，拡張の方向性としては，

知名度に関する指標の拡張　知名度の表現にフォロワの数だけでなく他の指標を取り入れるという拡張

選挙地盤に関する指標の追加　選挙地盤を表現するような指標を新たに提案し取り入れるという拡張

の 2 つの拡張が考えられる．本研究では，第 1 段階として知名度に関する指標の拡張を行い，第 2 段階として選挙地盤に関する指標の追加を行うことで精度の向上を目指す．

3.1 知名度に関する指標の拡張

Cameron らの研究は知名度を表現する指標としてフォロワの数のみを利用していたが，Twitter のデータから得られる知名度を表現する可能性のある指標はフォロワの数だけではない．たとえば，知名度を表現する可能性のある指標とそのアイディアとして下記のものが考えられる．

フレンド数　多くのユーザとフレンドになることで，知名度を高められる可能性がある．一方で，Twitter では歌手やタレントなど人気のあるユーザはしばしばフォロワ数が大きい一方で，フレンド数が非常に小さいということがあるため，フレンド数が小さいと知名度が高い可能性がある．いずれにせよ，フレンド数は知名度を表現する指標である可能性が高い．

被登録リスト数　候補者が登録されているリストの数を非登録リスト数と呼ぶこととする．候補者をリストに登録するということは他のユーザとは分けてツイートを受け取るということであり，登録されているリストの数が大きいと知名度が高い可能性がある．

認証バッジの有無　Twitter がそのアカウントが本人のものであると確認したという証拠のマークである認証バッジは芸能人やスポーツ選手，政治家などの著名な人のアカウントを中心に付けられている [14] ため，認証バッジのあるアカウントは知名度が高い可能性がある．

存在日数　Twitter を長く利用することで，Twitter における知名度が高まる可能性がある．

選挙期間中のツイートの数　活発にツイートし有権者に発信することで候補者の知名度が高まる可能性がある．

上記の指標の中には選挙地盤を表現する可能性のある指標もあると考えられるが，知名度と選挙地盤には関連があり厳密に分類することが難しいため，ここでは特に知名度を表現する指標として扱うこととする．他にもこのような指標があると考えられるが，本研究ではフォロワ数に加え上記の 5 つの指標を利用することで知名度の表現を拡張する．以降では，これらの 6 つの指標を**拡張した知名度指標**と呼ぶこととする．

3.2 選挙地盤に関する指標の追加

ここでは，先に述べたように，選挙地盤を表現するような指標を新たに提案する．まず，Twitter 上で選挙地盤を構成する支持者や後援会に相当するユーザのアイディア，および選挙地盤を表現するような指標のアイディアを述べ，次に，その指標を定式化する．

3.2.1 選挙地盤に関する指標のアイディア

支持者や後援会は候補者が選挙でより多くの票を獲得し

当選するように候補者の選挙運動を支援するが，支持者や後援会はTwitter上ではどのようなユーザだろうか．おそらく，Twitter上では候補者のツイートをよくリツイートしているユーザであると考えられる．なぜなら，候補者がTwitter上でより多くの有権者に働きかけるためツイートによる情報発信を行うなかで，自分のフォロワに他のユーザのツイートを再投稿する機能であるリツイートは，より多くの有権者に候補者のツイートが届くように支援する方法として最も有用な方法のうちの1つだと考えられるからである．必ずしもすべての支持者や後援会がリツイートを通して候補者を支援するわけではなく，また他にもTwitter上で候補者を支援する方法はあると考えられるが，本研究では，特にリツイートにより候補者の情報発信を支援するユーザをTwitter上での支持者や後援会とし，以降ではこのような支持者や後援会のユーザを単に**後援者**と呼ぶこととする．

次に，選挙地盤を表現するような指標のアイディアについて述べる．選挙地盤を表現するような指標はいくつか考えられ，たとえば下記のものがある．

選挙地盤のリーチ　後援者はTwitter上で候補者のツイートをリツイートすることで候補者の選挙運動を支援するため，選挙地盤を構成する後援者が多ければ候補者に有利であり，したがって後援者の数は候補者の選挙地盤と当選の関係をよく表現している可能性がある．しかし，Twitter上の後援者は規模の大きい後援団体のアカウントや規模の小さい個人のアカウントなどがあり多様であると考えられ，候補者のツイートをリツイートにより広める頻度や1リツイートあたりに広めるユーザの数が異なるため，それらを考慮した方が候補者の選挙地盤と当選の関係をよく表現できると考えられる．ここでは，選挙地盤を構成するすべての後援者が候補者のツイートを広めるユーザの数の期待値を**選挙地盤のリーチ**と呼ぶこととする．選挙地盤のリーチは後援者の数では表現しきれていない候補者の選挙地盤と当選の関係をよく表現できると考えられる．

選挙地盤のバラエティ　選挙地盤を構成する後援者の多様さ（相互に結合されていないか）を**選挙地盤のバラエティ**と呼ぶこととする．選挙地盤のバラエティは選挙地盤のリーチで表現しきれていない選挙地盤と当選の関係をよく表現している可能性がある．2人の候補者XとYを考え，XとYは選挙地盤のリーチが等しいとする．このXとYは，ほぼ同数のユーザにツイートを届けることができるが，X，Yはそれぞれの後援者同士のつながりには違いがある．Xは，後援者が完全に結合しており，Xの後援者同士は相互フォローしている．Yは，後援者が完全に独立しており，どの2人の間にもフォローの関係がない．このとき，Xの後援者とYの後援者のどちらが候補者にとって力になるだろうか．答えはおそらくYである．なぜなら，後援者同士が相互フォローしていないということは，後援者同士の関係が薄い可能性が高く，したがって異なるコミュニティに所属するユーザである可能性が高い．実世界ではそれぞれの後援団体がそれぞれ多くの有権者をかかえていることが多いことと，Twitter上で候補者を大規模に支援する後援者が後援団体のアカウントである場合があることを考慮すると，後援団体がかかえている有権者が重ならないよう多様な後援団体から支援を受けている候補者の方が実世界の得票という点で有利である可能性がある．なお，Conitzerの研究では，政治における選挙であるかどうかにかかわらず，より密に結合したソーシャルネットワークを持つユーザは選挙において集団投票を行いやすいことを示唆しているが[13]，特に本研究では，政治における選挙を考慮しているという点で後援者の多様性の方が重要であると考える．

選挙地盤のロイヤルティ　選挙地盤を構成する後援者の忠誠度を**選挙地盤のロイヤルティ**と呼ぶこととする．選挙地盤のロイヤルティは選挙地盤のリーチやバラエティで表現しきれていない選挙地盤と当選の関係をよく表現している可能性がある．2人の候補者XとYを考え，XとYは選挙地盤のリーチもバラエティも等しいとする．Xの後援者は，その候補者だけでなく他の候補者もフォローしており，よくリツイートする．一方，Yの後援者は，（選挙の候補者の中では）その候補者だけをフォローしており，他の候補者のリツイートはいっさいしない．このとき，Xの後援者とYの後援者のどちらが候補者にとって力になるだろうか．答えはおそらくYである．候補者自身に関してのリツイートを広める力は同じでも，他の候補者のツイートを流すことで，他の候補者の認知も上げてしまうからである．得票率の向上という点で，候補者にとっては情報拡散は排他的である方が良い．

このほかにも，たとえば，後援者のツイートの内容（どのくらいポジティブか，どのくらい選挙について語っているか），リツイートの時間的な局在性（他の後援者がリツイートしていない日時にリツイートしているか），リアルなコミュニティへの影響（どのくらい社会的な地位のあるユーザか）など，さまざまな拡張が考えられるが，本研究では，Twitterユーザのつながりだけから定義することのできる上記の3つの指標に絞って考え，また，以降では，これらの指標を**選挙地盤指標**と呼ぶこととする．

3.2.2 指標の定式化

まず，後援者の定式化について述べる．候補者cの期間中のツイートの集合をT_c，cのツイートをリツイートしたユーザをu_i（$u_i \in F_c \cup \{c\}$，ただし，F_cはcのフォロワ集合），cのツイートの中でu_iがリツイートしたツイートの

集合を $t_{c,i}$ とすれば，c のツイート全体における u_i がリツイートしたツイートの割合は

$$\alpha_{c,i} = |t_{c,i}|/|T_c|$$

となる（ただし，u_i が c の場合は情報拡散はリツイートではなくツイートにより行われるが，ここでは，説明の便宜上 u_i が c の場合もツイートではなくリツイートと表記することとする．また，u_i が c の場合は $\alpha_{c,i}$ は 1 とする）．リツイートすると，u_i のフォロワ F_{u_i} に広がるため，u_i が候補者 c の 1 ツイートを拡散するユーザ数の期待値 $r_{c,i}$ は

$$r_{c,i} = \alpha_{c,i} \times |F_{u_i}|$$

となる．ここで，$r_{c,i} \geq r_{thre}$ を満たす $u_i \in F_c$ を c の後援者と定義し，U_c とする．後援者はよくリツイートすることで多くのユーザに候補者のツイートを広げるユーザであることと，$r_{c,i}$ が小さいフォロワを加えても reach の値が大きく変わらず，また，処理時間は $r_{c,i}$ の多寡にかかわらず同じようにかかることを考慮して，閾値を定める．予備実験で，指標の計算結果と計算時間のトレードオフを考慮し，$r_{thre} = 100$ とした．すなわち，候補者のツイート 1 つにつき，平均 100 ユーザ以上に広めているユーザが候補者 c の後援者である．

次に，選挙地盤指標の定式化について説明する．第 1 に，選挙地盤のリーチについて，候補者 c の後援者 $u_i \in U_c$ の全フォロワ集合を $G_c = \bigcup_{u_i \in U_c} F_{u_i}$ とし，u_i が $g_j \in G_c$ の g_j に候補者 c のツイートを広める割合を $s_{i,j}$ とし，選挙地盤のリーチは次のように定義する．

$$\text{reach}_c = \sum_{g_j \in G_c} (1 - \prod_{u_i \in U_c} (1 - s_{i,j}))$$

すなわち，候補者の 1 ツイートが拡散されるユーザ数の期待値である．

第 2 に選挙地盤のバラエティについて，候補者 c の後援者 $u_i \in U_c$ が，$u_j \in U_c$（ただし $i \neq j$）と相互にフォローしている関係でない割合を $v_{c,i}$ とし，選挙地盤のバラエティは次のように定義する．

$$\text{variety}_c = \frac{\sum_{u_i \in U_c} r_{c,i} \times v_{c,i}}{\sum_{u_i \in U_c} r_{c,i}}$$

すなわち，リーチで重みづけた後援者ごとの，相互フォローでない割合の平均である．なお u_i と u_j が相互フォローしている関係であるとは，u_i が u_j をフォローしており，同時に u_j が u_i をフォローしていることをいう．

第 3 に選挙地盤のロイヤルティについて後援者 $u_i \in U_c$ がリツイートする全候補者のツイートに対する候補者 c のツイートの割合を $l_{c,i}$ とし，選挙地盤のロイヤルティを以下のように定義する（ただし，u_i が c の場合は，$l_{c,i}$ は 1 とする）．

$$\text{loyalty}_c = \frac{\sum_{u_i \in U_c} r_{c,i} \times l_{c,i}}{\sum_{u_i \in U_c} r_{c,i}}$$

すなわち，リーチで重みづけた後援者ごとの，リツイートにおける候補者 c のツイートの割合の平均である．

以上，選挙地盤指標の定式化について述べた．より複雑な定式化もありうるかもしれないが，いずれも，指標のもともとのアイディアをシンプルに定式化したものである．

本章では，Cameron らの手法の拡張として知名度に関する指標の拡張と選挙地盤に関する指標の追加の 2 つの拡張についてアイディアとその定式化について説明した．次章では，本研究で予測実験に用いるデータについて説明する．

4. データ

本章では，本研究で予測実験に用いる 2013 年参議院議員選挙の結果とその選挙期間における Twitter のデータについて説明する．まずデータの取得方法および取得結果について概説し，次に簡単な分析を通してデータの概観を示す．

4.1 データの取得

2013 年の参議院議員選挙を対象としてデータの収集を試みた．当該選挙の候補者は 433 人であり，そのうちインターネットを活用した選挙運動を行える期間（以下，選挙期間）に Twitter を利用していた候補者は 287 人であった．候補者のうち 77 人が当選で 210 人が落選であった．

Twitter からのデータ取得について，知名度に関する指標の算出のために，アカウントのプロフィールに関するデータを，選挙期間開始時と選挙期間終了時の 2 時点で取得した．また，選挙地盤に関する指標を算出するために，候補者のツイートと候補者のツイートのリツイートをすべて取得した．候補者のツイートは 42,645 件得られ，また候補者のツイートのリツイートは 368,694 件得られた．リツイートのデータは，まず検索 API を用いて候補者アカウントのスクリーン名を含むツイートを取得し，次に取得したツイートの中から候補者のツイートへのリツイートを抽出し取得した．なお，以後の評価実験では実際の候補者の当落とあわせた学習を行っているが，今回使用したデータはいずれも選挙期間中のものであり，選挙の当落が決まったあとのデータは含まれていない．また，これらのデータは Twitter の REST API を用いて取得した[*1]．

4.2 データの概観

Twitter から取得したデータの概観を示す．まず，各候補者のフォロワ数と期間中のツイート数，ツイートがリツイートされた回数の関係を**図 1** に示す．フォロワ数とリツイートされた数の相関係数およびツイート数とリツイートされた数の相関係数はそれぞれ 0.283，0.312 と小さく，必ずしもフォロワ数や期間中のツイート数が大きければ，リツイートによる情報拡散を期待できるわけではないことが

[*1] 共同研究により特別な API を用いた．

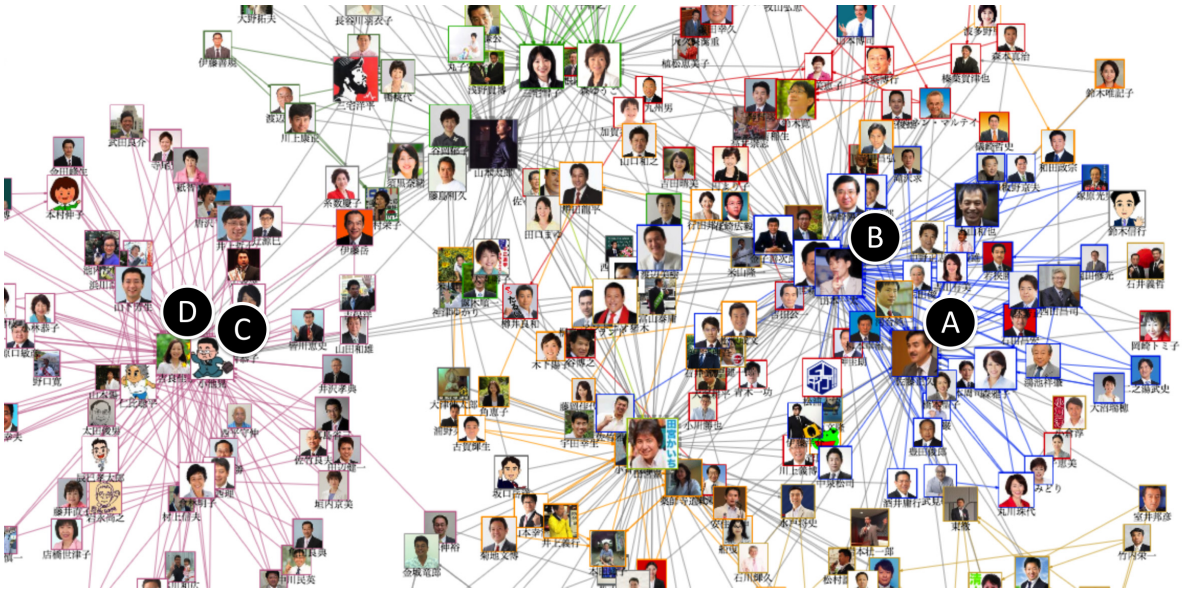

図 2 フォロワ集合の類似度に基づいて構築した候補者ネットワーク．記号は本文中で言及する候補者の右上に付加している

Fig. 2 The candidates' followers cooccurence network.

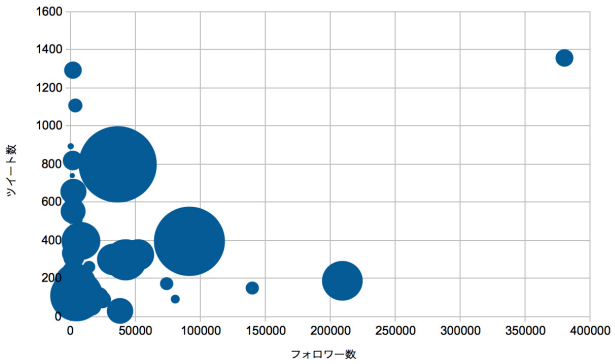

図 1 各候補者のフォロワ数と選挙期間中のツイート数とツイートをリツイートされた回数の関係．バブルの大きさは候補者の被リツイート数を表す

Fig. 1 The relation among the number of tweets, the number of followers and the number of times in which tweets are retweeted for each candidate. The size of bubbles represents the retweeted count.

分かる．

次に，候補者間の後援者による関係を俯瞰的に把握するために，候補者のフォロワ集合の共起からフォロワの類似度を評価し，候補者の投稿を受け取る有権者の重なりを可視化する．具体的には，2 人の候補者 X，Y のフォロワ集合の大きさを $|X|$，$|Y|$，AND 集合の大きさを $|X \cap Y|$ としたときに，下記の式により定義される Simpson 係数によりフォロワの類似度を評価する．

$$\text{Simpson}(X,Y) = \frac{|X \cap Y|}{min(|X|,|Y|)} \quad (1)$$

全候補者ペアについて Simpson 係数を算出し，各候補者から Simpson 係数が高い上位 3 人の候補者に対してエッジを引くことでネットワークを構築し可視化した．可視化したネットワークを図 2 に示す．可視化に際して，同政党の候補者は同色枠のノードで表し，ノードの大きさはフォロワ数の対数に比例するように定めた．青は自民党で，赤は民主党，紫は共産党である．自民党は A) 佐藤正久氏や B) 山本一太氏を中心として密なクラスタを形成している．一方で，共産党は C) 小池晃氏や D) 吉良佳子氏を中心として密なクラスタを形成しているが，共産党の候補者は他政党の候補者との間にエッジが少なく，したがって，共産党の候補者のフォロワは他政党の候補者をフォローしない傾向にあると推察される．民主党は中心となる候補者がおらず，粗なクラスタとなった．可視化したネットワークについて，約 60%のエッジは同政党の候補者間にあり，候補者の投稿は所属政党に興味関心がある有権者に対して，より拡散されやすい状況であったと考えられる．

5. 評価実験

本章では，Cameron らの手法に加えた 2 つの拡張，特に選挙地盤に関する指標の追加により予測精度が改善されるかを検証するため評価実験を行う．まず，実験の設定について説明し，次に実験結果について述べる．

5.1 実験設定

本評価実験では 2013 年の参議院議員選挙の当選者を予測するモデルを教師あり学習のアルゴリズムより構築し，そのモデルの予測精度の評価を行う．当該選挙に出馬し選挙期間中に Twitter を利用していた候補者は 287 人で，当選者は 77 人であり本実験では当選を正例とする．教師あり学習のアルゴリズムはナイーブベイズや SVM などさまざまなアルゴリズムがあるが，ここでは学習後に各素性の

重みを確認でき，また広く用いられ良好な結果が得られている Random Forest を用いる．予測精度の評価は 10 分割交差検定で行い，評価指標には正解率と精度，再現率，F 値の 4 指標を利用し，特に F 値で予測モデルを評価する．

モデルの予測精度の比較では，1) ランダムに予測する手法，2) Cameron らの手法，3) 拡張した知名度指標を素性とする手法，4) 拡張した知名度指標と選挙地盤指標を素性とする手法の 4 手法を比較する．1) ランダムに予測する手法は 77/287 の確率で当選と判定し 210/287 の確率で落選と判定する予測モデルであり，ベースラインとして設ける．2) Cameron らの手法は 4 時点でのフォロワの数を素性とするものであったが，4 時点の指標を用いてもほとんど精度が得られなかったことと本評価実験の対象の選挙では選挙期間の少し前から Twitter を利用し始めた候補者もいたことを考慮して選挙期間開始時と選挙投票前日のフォロワ数の 2 指標を素性とする条件でも十分であると考えこの条件下で利用した．3) 拡張した知名度指標を素性とする手法は，選挙投票前日のフォロワ数，フレンド数，被登録リスト数，認証バッジの有無，存在日数，選挙期間中のツイート数の 6 つの指標を素性とする手法である．4) 拡張した知名度指標と選挙地盤指標を素性とする手法は 3) で用いる 6 つの指標に加えて選挙地盤のリーチ，バラエティ，ロイヤルティの 3 つの選挙地盤指標を素性とする手法である．

なお，本評価実験では，過去の選挙結果などから学習し選挙結果を事前に予測するのではなく，Cameron らの研究 [8] と同じく，同一の選挙の別の候補者の結果を用いてモデルを学習し，その他の候補者の当落を予想しているので，通常の意味での予測ではない．交差検定を行うことで，学習モデルにおける素性の寄与を見ることが目的であり，そうして得られた素性は，過去の選挙と照らし合わせることで今後の選挙の分析や予測に活用できる可能性がある．

5.2 実験結果

まず，4 つの手法の予測精度の結果を**表 1** に示す．これらの結果を比較することで得られた 4 つの主要な知見を下記に述べる．第 1 に，Cameron らの手法はランダム予測より F 値が高く，確かにフォロワ数は選挙予測に有効であることが分かったが，一方で，それほど高い精度が得られていないことから，フォロワの数だけを考慮しても選挙結果を予測できないとする Cameron らの結論と合致する結果となった．第 2 に，拡張した知名度指標を素性とする手法は Cameron らの手法より F 値が高く，このことはフォロワ数だけではなく他の指標も知名度を表現している可能性があることを示唆している．第 3 に，拡張した知名度指標と選挙地盤指標を同時に素性とする手法は拡張した知名度指標のみを素性とする手法よりも F 値が約 12%高く，このことは本研究で提案した選挙地盤指標が選挙当落の予測に有効であることを示し，また，本研究で提案した選挙地盤指標が古くから選挙当落で重視されてきた選挙地盤をよく表現している可能性があることを示唆している．第 4 に，拡張した知名度指標と選挙地盤指標を同時に素性とする手法は Cameron らの手法と比べると F 値は約 70%高く，このことは，Cameron らが示唆する僅差で競っているケースだけでなく小さいもののソーシャルメディアが選挙当落に影響を与えるケースがある可能性があることを示唆している．

次に，拡張した知名度指標と選挙地盤指標を素性とする手法の実験の際に構築した予測モデルの各素性の重みを**表 2** に示す．素性の重みとともに，各素性と選挙当落（当選を 1，落選を 0）への相関分析を行い，相関係数もあわせて記載した（Random Forest より得られる素性の重みはすべて正であり，当落のどちらに寄与しているか分からないため，符号は相関係数の符号に合わせて記載している）．素性の重みと相関係数の絶対値の大小はおおむね一致している．

まず，拡張した知名度指標のそれぞれの重みから得られる知見や重みに対する考察を下記に述べる．

- フォロワ数
  Cameron らの手法でも利用されていたフォロワ数の重みは大きく確かに有効な素性であった．
- フレンド数
  フレンド数が負の重みになっているのは一見不思議であるが，Twitter では歌手やタレントなど人気のある

表 1 10 分割交差検定による予測モデルの評価
Table 1 Prediction result with 10-fold cross validation.

利用指標または手法	正解率	精度	再現率	F 値
ランダム予測	0.607	0.268	0.268	0.268
Cameron らの手法	0.702	0.508	0.280	0.335
拡張した知名度指標	0.766	0.573	0.455	0.507
拡張した知名度指標＋選挙地盤指標	**0.780**	**0.658**	**0.499**	**0.568**

表 2 拡張した知名度指標と選挙地盤指標の重み，および選挙当落との相関係数
Table 2 The feature weights and correlations to the election result in the prediction using two types of features.

カテゴリ	素性	重み	相関係数
拡張した知名度指標	フォロワ数	**0.102**	0.124
	フレンド数	**−0.235**	−0.0376
	被登録リスト数	**0.242**	0.236
	認証バッジの有無	0.00154	0.0563
	存在日数	0.0790	0.0383
	選挙期間中のツイート数	−0.0838	−0.0632
選挙地盤指標	選挙地盤のリーチ	**0.100**	0.114
	選挙地盤のバラエティ	0.0592	0.0815
	選挙地盤のロイヤルティ	**0.0970**	0.113

ユーザはしばしばフォロワ数が大きい一方で，フレンド数が非常に小さいということがある．逆に，有名でない候補者の方がフレンド数は多いことがあるために，このような結果になっていると考えられる．特に本実験では対象が国政選挙の候補者であり少なくないユーザが著名な人であったため，フレンド数が知名度をよく表現できたのかもしれない．また，有権者をフォローすることで知名度が高まる可能性も考えられたが，重みが負であることからフォローすることの効果は大きくないのかもしれない．

- 被登録リスト数
  被登録リスト数の重みは最も大きく，候補者をリストに登録するということは他のユーザとは分けてツイートを受け取るということであり，そのような熱心なユーザに関心を持たれる方が当選しやすいということが推察される．
- 認証バッジの有無
  認証バッジの有無は重みの絶対値が最も小さく，選挙の当落とはほとんど関係がなかったといえる．認証バッジは著名な人のアカウントを中心に付けられているため，知名度が高い可能性があると考えたが，Twitter が認める著名な人であることと選挙当落で重視される知名度が高いということは，必ずしも合致しているわけではないことが推察される．
- 存在日数
  存在日数の重みは正であまり大きくなかったがTwitterを長く使っている候補者はその分 Twitter における知名度が高いという仮説と矛盾しなかった．
- 選挙期間中のツイート数
  選挙期間中のツイート数の重みは負であった．選挙期間中のツイートの多さが落選に寄与するという解釈もできるが，もともと当選しにくい候補者が活発にツイートすることで挽回の機会を狙っていた可能性がある．本評価実験の対象データである 2013 年の参議院議員選挙はインターネットの利用が解禁された初めての国政選挙ということもあり，当選への新しい可能性を見い出すためもともと当選しにくい一部の候補者はインターネットを積極的に活用していたと推察される．

次に，選挙地盤指標のそれぞれの重みから得られる知見や重みに対する考察を述べる．選挙地盤のリーチやロイヤルティが，重み，相関係数ともにフォロワ数と同程度に高く，これらの指標は候補者の状態と当選の関係をよくとらえていると考えられる．一方で，選挙地盤のバラエティの重みはそれらと比較して小さかった．この指標は多様な後援団体から支援されている方が有利であるというアイディアに基づいたものであることをふまえると，重みが小さかった原因として，本指標では実世界の後援団体の多様度をうまく表現できていない可能性や Conitzer の研究で示唆されるようにより密に結合したソーシャルネットワークを持つ候補者は後援者や支持者は集団投票を行いやすいということの影響が政治の選挙でも小さくなかったという可能性が考えられる．また，実験全体を通して選挙地盤指標が選挙結果の予測に有効であったことと，特に，選挙地盤のリーチとロイヤルティの予測への寄与が高かったことをふまえると，地盤が強固であるということは，Twitter において候補者やその後援者が多くのユーザに排他的にツイートを広めることができる状態なのかもしれない．

以上より，本評価実験からいえることは以下である．

- 従来手法で用いられいたフォロワ数は確かに有効な素性である．
- Twitter 特有の被登録リスト数は，フォロワ数よりも有効であり，またフレンド数も有効な素性である．基本的にこれらは候補者の知名度を表す指標であると考えられる．
- 一方で，本研究で提案した選挙地盤指標も十分に有用である．選挙地盤リーチや選挙地盤ロイヤルティは有効な素性であり，またそれらほどではないが選挙地盤のバラエティも有効な素性である．

次章では，これまでの本稿の内容をふまえて本研究の限界を述べ，課題と拡張可能性について考察する．

6. 考察

本章では，これまでの本稿の内容をふまえて，まず，本研究の限界を述べ，次に，予測精度を向上させるための課題と拡張可能性について考察する．

まず，本研究の限界について考察する．本研究は選挙地盤が選挙当落の重要な要素であるという日本の選挙の特徴に基づいている．したがって，本研究で提案した選挙地盤指標は，選挙地盤が重視される選挙に対しては有効であると考えられるが，選挙地盤が重視されない選挙に対しては有効でない可能性がある．また，特に日本では Twitter は日本語での利用が中心であり，他国のユーザが日本語で記載されたツイートをリツイートすることは多くないと考えられるが，他国の選挙で特に候補者が英語で情報発信するような場合では，注意が必要であると考えられる．なぜなら，候補者のツイートに関心を持ちリツイートするユーザが当該選挙の有権者や選挙に関与する団体である可能性は日本の場合と比較して低いと考えられるからである．したがって，本研究で提案した選挙地盤指標の他国の事例への適用可能性は限定的である可能性が高いと考えられる．

次に，予測精度の向上に向けた課題について考察する．まず，Twitter の利用率について，当該選挙において候補者 433 人のうち 287 人（66%）が，また当選者 121 人のうち 77 人（64%）が Twitter を利用した．これを高いと見るか低いと見るかは議論の分かれるところであるが，Twitter を使うことは手軽にできることを考えると，この数字はけっ

して高いとはいえないと考える．インターネット選挙運動が解禁された初めての国政選挙であったことや候補者の多くの方が中高年の方（候補者の平均年齢は 51 歳）であったことで全体の利用率が高くなかったのではないかと考えられる．中高年層よりは若年層の方がより Twitter を利用していたと考えられ，今後の国政選挙ではより多くの候補者の Twitter 利用が期待できると考えられる．したがって，本研究の適用可能性は今後大きくなるかもしれない．次に，選挙方式について，本研究の対象である参議院議員選挙では選挙方式が選挙区制と比例区制の 2 つがあり，選挙区制に出馬する候補者は出馬した選挙区内で他の候補者と票を競い，比例区制に出馬する候補者は比例区制に出馬する全国の候補者と票を競うことになっており，選挙方式がやや複雑である．単純に候補者の指標だけでなく，どの選挙区で出馬するかも重要であるため，適用可能性を向上させるためには，選挙区ごとで指標を相対化するなど工夫が重要であると考えられる．

最後に，本研究の拡張可能性について考察する．本研究の拡張は大きなものとして，1) 鞄の指標を取り入れるという拡張，2) 有権者に焦点を当てた指標を取り入れるという拡張，の 2 つがあると考えられる．1 つ目の鞄の指標を取り入れるという拡張について，本研究は三バンの中で特に地盤と看板の概念を扱ったものであるが，鞄の概念は扱わなかった．選挙資金の豊富さは Twitter のデータからの予測が難しいと考えられるため取り入れなかったが，鞄の概念は地盤と看板と同様に選挙当落に影響のある重要な要素であるため，取り入れることができれば，予測精度は改善される可能性が高い．2 つ目の有権者に焦点を当てた指標を取り入れるという拡張について，選挙結果の予測を試みる研究は候補者に焦点を当てる研究と有権者に焦点を当てる研究がある．有権者に焦点を当てる研究は主に政党名や候補者名に言及したツイートの分析を行うものである．選挙の当選者は候補者の中から有権者が投票により選定するものであり，両方の指標を取り入れることで有権者と候補者と当選の関係をよりよく表現できる可能性が高い．したがって，有権者に焦点を当てた指標を取り入れるという拡張は予測精度の向上を十分期待できる拡張であると考えられる．

7. まとめ

本研究では，選挙結果を高い精度で予測するモデルの構築を目指し，社会学で古くから選挙当落の重要な要素の 1 つとされてきた選挙地盤を定量的に測定，指標化し，この指標を用いることで既存研究の拡張を試みた．選挙地盤に関する指標は選挙地盤のリーチ，バラエティ，ロイヤルティという 3 つの指標を提案した．選挙運動へのインターネットの利用が初めて解禁された 2013 年の参議院議員選挙を対象とした評価実験の結果，本研究で提案した 3 つの選挙地盤指標は選挙結果の予測に有効であることが分かり，また，本研究で用いた手法は先行研究の手法と比較して F 値が約 70%高く，Twitter の利用は選挙結果に小さいものの影響があることが示唆された．実験全体を通して，選挙地盤が強固であるということは，Twitter において候補者やその後援者が排他的にツイートを広められるユーザの数が多い状態に相当することが示唆された．

今後も，インターネットを活用する有権者が増加するとともに，インターネットを活用した選挙の形は進化していくものと考えられる．本研究では，特に Twitter に焦点を当てて，古くから重要だとされてきた候補者の選挙地盤の強さを，ソーシャルメディアの上で測定する試みである．こうした測定手法を提案することで，従来からの社会学の研究に新しい切り口を与えることができるかもしれない．さらには，本研究が今後のインターネット選挙運動を活性化し，よりよい社会を構築する一助となれば筆者らの幸いとするところである．

参考文献

[1] 衣笠達夫：人口急増都市における政治的選択と財政支出の分析，地域学研究，Vol.11, pp.119–134 (1980).

[2] 荒牧英治，増川佐知子，森田瑞樹：Twitter Catches the Flu：事実性判定を用いたインフルエンザ流行予測，情報処理学会研究報告．SLP，Vol.2011, No.1, pp.1–8 (2011).

[3] Bollen, J., Mao, H. and Zeng, X.: Twitter mood predicts the stock market, *Journal of Computational Science*, Vol.2, No.1, pp.1–8 (2011).

[4] Tumasjan, A., Sprenger, T.O., Sandner, P.G. and Welpe, I.M.: Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment, *Proc. 4th International AAAI Conference on Weblogs and Social Media* (2010).

[5] Sang, E.T.K. and Bos, J.: Predicting the 2011 dutch senate election results with Twitter, *Proc. 13th Conference of the European Chapter of the Association for Computational Linguistics* (2012).

[6] Jungherr, A., Jurgens, P. and Schoen, H.: Why the Pirate Party Won the German Election of 2009 or The Trouble With Predictions: A Response to Tumasjan, A., Sprenger, T.O., Sander, P.G. and Welpe, I.M.: Predicting Elections With Twitter: What 140 Characters Reveal About Political Sentiment, *Social Science Computer Review*, Vol.30, No.2, pp.229–234 (2012).

[7] O'Connor, B., Balasubramanyan, R., Routledge, B.R. and Smith, N.A.: From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series, *Proc. 4th International AAAI Conference on Weblogs and Social Media* (2010).

[8] Cameron, M.P., Barrett, P. and Stewardson, B.: Can Social Media Predict Election Results? Evidence from New Zealand, *Working Paper in Economics*, Vol.13, No.08 (2013).

[9] Kwak, H., Lee, C., Park, H. and Moon, S.: What is Twitter, a Social Network or a News Media?, *Proc. WWW '10*, *Proc. 19th International Conference on World Wide Web* (2010).

[10] Akioka, S., Kato, N., Muraoka, Y. and Yamana, H.: Cross-media impact on twitter in japan, *Proc.*

SMUC '10, Proc. 2nd International Workshop on Search and Mining User-generated Contents (2010).
[11] Gayo-Avello, D., Metaxas, P.T. and Mustafaraj, E.: Limits of Electoral Predictions Using Twitter, *Proc. 5th International AAAI Conference on Weblogs and Social Media* (2011).
[12] Chung, J. and Mustafaraj, E.: Can collective sentiment expressed on twitter predict political elections?, *Proc. 24th AAAI Conference on Artificial Intelligence* (2011).
[13] Conitzer, V.: Should social network structure be taken into account in elections?, *Mathematical Social Sciences*, Vol.64, No.1, pp.100–102 (2012).
[14] Twitter JP：認証済みアカウントについて，Twitter（参照 2014-01-21）.

那須野 薫

2013 年東京大学工学部システム創成学科卒業．2015 年現在，同大学大学院工学系研究科技術経営戦略学専攻修士課程在籍．専門は，ソーシャルメディア分析，ビッグデータ分析，学習科学．

奥山 晶二郎

2000 年朝日新聞入社．佐賀支局等を経て 2007 年にデジタル部門へ．2012 年からデジタル編集部記者．SNS を活用した企画「ビリオメディア」や，データジャーナリズムの手法による震災報道等を担当．

中西 鏡子

WEB 制作会社等を経て，2012 年 5 月より朝日新聞社デジタル編集部クリエイティブ開発チームにて，朝日新聞デジタルのコンテンツ制作に従事．これまでに災害やツイッター分析関連コンテンツ等を制作．

松尾 豊 （正会員）

2002 年東京大学大学院博士課程修了．博士（工学）．産業技術総合研究所，スタンフォード大学を経て，東京大学准教授．人工知能学会編集委員長を経て，倫理委員長．専門は，ディープラーニング，Web 工学，人工知能．