フォローを用いた特定地域から発信されたツイートの効率的な収集

中川　真史† 　山口　祐人†† 　北川　博之†††

† 筑波大学システム情報工学研究科　〒 305–8573 茨城県つくば市天王台 1–1–1
†† Indeed Japan 株式会社　〒 100–0006 東京都千代田区有楽町 2–7–1 有楽町イトシア 15 階
† 筑波大学計算科学研究センター　〒 305–8573 茨城県つくば市天王台 1–1–1
E-mail: †nakagawa@kde.cs.tsukuba.ac.jp, ††yuto.ymgc@gmail.com, †††kitagawa@cs.tsukuba.ac.jp

あらまし　本研究では Twitter におけるユーザ間のフォロー情報を用いて，特定地域より発信されたツイートを効率的に収集する手法を提案する．特定地域に対するイベント検出，災害分析などの要求実現のためには，その地域から発信されたツイートの収集効率化が 1 つの重要な技術課題となる．目的ツイートの収集を効率的に行うために，その目的ツイートを多数発信するローカルユーザをフォローするアプローチが効率的である．このアプローチ方法は新規フォローユーザの探索アルゴリズムやツイートの発信地推定といった技術要素で構成される．本研究は既存手法による新規フォローユーザ探索の方法が効率的でないという課題に取り組んだ．ユーザ間のフォロー情報を新規フォローユーザの探索アルゴリズムへ導入することで，特定地域から発信されたツイートの収集を効率化する．そして実データを用いた評価実験により，提案手法の有効性を示す．

キーワード　Twitter，Focused Crawling，Bandit Problem with Side Information

1. はじめに

多数のユーザがスマートフォンから SNS（ソーシャルネットワーキングサービス）を通して身の回りの出来事を含む多様な情報を発信している．Twitter ではツイートを携帯端末から手軽に投稿できるため，地域と結びついたコンテンツが発信されている．そのため，イベント検出 [1] や災害分析 [2] など，地域と結びついたツイートコンテンツの利用が要求されている．その要求実現のためには，特定地域から発信されたツイートの収集効率化が 1 つの重要な技術課題となる．

しかしながら，Twitter 社が提供する既存のツイート収集 API(注1) の枠組みの中で特定地域から発信されたツイートを収集することは計算機的・時間的効率が悪い．収集が低効率である理由は，Twitter API には厳しい利用制限があることと，ツイートの発信地を示すジオタグ付きのツイートの割合が少ないことにある．リアルタイムでツイートを収集する際には Twitter Streaming API を用いる． Streaming API では取得可能ツイート数が全体に占める割合に制限がかけられている．例えば Streaming API の中の statuses/sample はリアルタイムでツイートを収集する API であるが，全ツイート中の 1%のランダムサンプルのみが収集可能である．また過去のツイートを収集するためには Twitter REST API を用いる． REST API の場合も利用回数と 1 回で取得できるツイート数に制限がかけられている．さらにジオタグ付きツイートを地理条件で検索可能な API もあるが，ツイートの発信地を示すジオタグ付きのツイートの割合が約 2% [3] と少ない．このように Twitter API の利用制限とジオタグ付きのツイート数の割合が少なさといった理由から Twitter API の枠組みでは取得可能なツイート数の割合が少なく，特定地域におけるトレンドやイベントを分析する際に十分な量のツイートを収集することは困難である．

先行研究 [4] では特定地域から発信されたツイートを収集するために，その地域からツイートを発信する可能性が高いローカルユーザをフォローする．ローカルユーザを発見するために新規フォローユーザの探索を行うと同時に，既知のローカルユーザをフォローし続ける．しかしながら，この既存手法は新規フォローユーザをランダム探索する点が非効率である．探索が常にランダムであるため，収集時間の経過とともに探索効率が向上しない．

そこで本研究は Twitter においてフォロー関係のあるユーザ同士は現実世界においても行動範囲が近い可能性があるという前提のもと，ユーザのフォロー情報を用いることで，新規フォローユーザの探索を改善し特定地域から発信されたツイートの収集の効率化を行う．

以降では，第 2 章で関連研究を述べた後，第 3 章では本研究が取り組む問題を定義し，第 4 章ではその問題に取り組んだ先行研究を説明する．続く第 5 章で先行研究の課題を解決する手法を提案し，第 6 章で評価実験の内容と結果について説明する．最後に第 7 章で本研究のまとめと今後の課題について述べる．

2. 関連研究

第 2.1 節でバンディット問題とその著名なアルゴリズムについて，第 2.2 節でバンディットアルゴリズムを用いたツイート収集手法について，第 2.3 節で side information を用いたバンディットアルゴリズムについて，第 2.4 節でツイートの発信地

(注1)：https://developer.twitter.com

やユーザの居住地推定手法について述べる．

2.1 バンディット問題・アルゴリズム

バンディット問題 [5], [6] とは，報酬の分布が異なる複数の選択肢があり，得られる報酬の期待値を最大化するような選択順序を求める問題である．報酬最大化のためには，報酬が高い選択肢を探索する必要があると同時に，探索によって発見した報酬が高い選択肢を可能な限り多く選択する（探索によって獲得した知識を活用する）必要がある．バンディットアルゴリズムはこのような探索と活用のトレードオフのもとで選択行動を最適化する．

解法が最もシンプルなバンディットアルゴリズムの一つとして ϵ-greedy アルゴリズム [7] が挙げられる．ϵ-greedy は確率 ϵ で報酬とは無関係に選択肢を探索し，確率 $1-\epsilon$ で過去の報酬が最大の選択肢を活用する．UCB(Upper Confidence Bound) アルゴリズム [6] は各選択肢の報酬と選択回数から，人為的に設定した有意水準によって，その選択肢の報酬の期待値の上限を算出し，その値が最大の選択肢を選ぶ手法である．この手法は選択回数が少ない選択肢と過去の報酬が高い選択肢を選択していき，低報酬な選択肢の探索を選択回数とともに減少させることで報酬最大化を行う．Thompson sampling アルゴリズム [8] は，各選択肢に対応した報酬確率分布からサンプリングした値が最大の選択肢を選択する手法である．各選択肢に対応した報酬確率分布を共通の事前確率分布で仮定しておき，得られた報酬によって事後確率分布を更新していく．この事後確率分布からサンプリングした値が最大の選択肢を選択していくことで報酬最大化を行う．

2.2 バンディットアルゴリズムを用いた特定トピックに関するツイートの収集

Gisselbrecht ら [9] は特定トピックについて言及したツイートを収集するために，そのトピックについてのツイートを多数発信するユーザをフォローする手法を提案した．フォローユーザが発信したツイート中の単語と特定トピックとの関連度（報酬）を計算し，バンディットアルゴリズムがその報酬に応じてフォローユーザを切り替える手法である．しかしながら，本研究の場合はツイート中の単語からそのツイートの発信地を直接算出することができない．そのため報酬を計算するために本研究では発信地推定を行う．本研究と Gisselbrecht らの手法ではフォローユーザの報酬算出方法が異なる．

2.3 Side information を用いたバンディットアルゴリズム

バンディット問題の実際的な例では，選択肢の報酬以外の情報（Side information）[10] が得られる場合がある．例えば，Web ページ訪問者に提示する広告選択問題の場合 [11]，提示する広告を選択する前に，ページ訪問者の性別・年齢層・地域や訪問サイト履歴，広告のトピックなどの情報が Side information にあたる．

バンディットアルゴリズムに Side information を導入することで，アルゴリズムの性能向上を実現した研究が複数行われている．LinUCB [12] はページ訪問者と広告の情報がベクトル化されているという前提のもと，そのベクトル値と過去に得られた報酬の関係を学習する．ベクトルを説明変数，報酬を従属変数とした線形回帰式のパラメタの学習をし，ベクトルから予想される報酬をもとに掲載広告を選択する．本研究は選択肢が複数の属性からなるベクトルを持つことを想定せずに選択肢同士のグラフ関係を用いる点でこの研究とは異なる．また Buccapatnam ら [13] は選択肢をノード，エッジをノード同士の任意の関係としたネットワークを構築し，ノードのネットワークにおける位置からそのノードを探索する確率・回数を最適化する手法を提案した．この Buccapatnam らの手法は，前もってネットワーク全体の情報を取得していることを前提としている．しかし，本研究では膨大な量の情報から目的の情報のみを取得するアプローチが要求される．本研究はある時点で報酬が高い選択肢と別の選択肢の関係を動的に取得し利用する探索手法を提案する．

2.4 ツイートの発信地やユーザの居住地推定手法

イベント検出や災害警報などの Twitter 利用のために，ツイートの発信地や Twitter ユーザの居住地の情報が必要となる．しかし，多数のツイートにジオタグが付いていないこと，多数のユーザは自らのプロフィールへ居住地を十分詳細に記入していないため，ツイートの発信地やユーザの居住地を推定する手法が多数提案されてきた．

個々のツイートの発信地を推定するために，Ikawa ら [14] は，ツイート中のキーワードとその発信地の関連付けを行う発信地推定手法を提案した．またツイート本文以外の情報を用いた研究として，Davis ら [15] はユーザ間のフォロー関係を用いたグラフベース手法，Schulz ら [16] はテキスト中の場所名やユーザプロフィールなど複数の情報を考慮したマルチチャネル手法によってツイートの発信地推定を行った．

Cheng ら [17] はツイート中から抽出した単語とその発信者の居住地の関係を学習した．ツイート中から単語を抽出する際，地理的に狭い範囲で多用されるローカルワードのみをツイート中から抽出することで推定性能を向上させた．Sadilek ら [18] は Twitter 上で交友関係にあるユーザ同士は地理的に近い場所にいるという前提のもと，ユーザの居住地を，Twitter 上で交友関係にある別のユーザのジオタグ情報を用いて推定する手法を提案した．他に Li ら [19] はツイートと交友関係の両方を用いてユーザの居住地を推定した．

これらの手法はツイートの発信地やユーザの居住地を推定することを目的としている．これらの研究と異なり，本研究はある特定の地域から発信されたツイートを収集することを目的としている．

3. 問題定義

本章では特定地域から発信されたツイートの収集タスクと以降の章で使用する記号の定義をする．本研究は先行研究 [4] を参考とし問題定義を行った．

まずツイートの収集期間を固定の一定時間幅のタイムウィンドウに区切る．タイムウィンドウ t におけるユーザ u の発信地 l からの収集ツイートを $tw = (u, t, l)$ と定義する．そして収集したツイート tw の集合を TW と定義する．TW は発信地が公開されているツイート集合 TW_{public} と公開されていないツ

表 1: 記号と意味

記号	定義
t	タイムウィンドウ
T	全タイムウィンドウ数
u	ユーザ
U	ユーザ集合
U_t	タイムウィンドウ t におけるフォローユーザ集合
U_{follow}	フォローユーザ集合 U_t の系列
tw	ツイート
TW	ツイート集合
TW_{public}	発信地が公開されているツイート集合
$TW_{private}$	発信地が公開されていないツイート集合
l	ツイートの発信地
L	発信地集合
$\hat{l}$	収集対象地域

イート集合 $TW_{private}$ へ分けられる．発信地が公開されているツイート集合 TW_{public} の発信地を l，その集合を L と定義する．

収集手法は，タイムウィンドウ T 終了後における収集対象地域 $\hat{l}$ から発信されたツイートの収集数 $|\{tw_i \in TW_{public} \mid l_i = \hat{l}\}| + |\{tw_i \in TW_{private} \mid l_i = \hat{l}\}|$ が最大となるように期待されるフォローユーザ集合の系列 $U_{follow} = \{U^1, U^2, ..., U^T\}$ とそれらフォローユーザから収集したツイートを出力する．本章で定義した記号とその意味を表 1 にまとめる．

4. 先行研究

上田ら [4] はバンディットアルゴリズムを用いて特定地域から発信されたツイートを収集する手法を提案した．目的ツイートを効率的に収集するために，その地域から発信するツイート数の期待値が高いローカルユーザをフォローするアプローチを取った．ユーザの選択においては，どのユーザがローカルユーザに該当するかという知識をあらかじめ持たない環境下においてローカルユーザを発見・フォローするために，バンディットアルゴリズムの 1 つである ϵ-greedy アルゴリズムが用いられた．この ϵ-greedy ツイート収集手法（以降，ϵ-greedy 手法）のタイムウィンドウ t における処理の流れは以下の 4 ステップである．

（1） タイムウィンドウ t の開始時:フォロー候補ユーザ集合 U から K 人を選択してフォローユーザ集合 U_t を決定する．

（2） タイムウィンドウ t の間:フォローユーザ集合が発信するツイートを収集する．

（3） タイムウィンドウ t の終了時:フォローユーザの評価（報酬計算）を行う．

（4） タイムウィンドウ $t+1$ へシフトしステップ 1. へ戻る．

上記 1. において，確率 ϵ で報酬とは無関係にランダムで 1 ユーザを選択し（新規ユーザの探索），確率 $1-\epsilon$ で過去の報酬が最大の 1 ユーザを選択するという方法を K 回繰り返すことでフォローユーザ集合を決定した．

上記 3 において，各ユーザをフォローした場合に得られる報酬を算出する．あるユーザの報酬をそのユーザが特定地域から発信するツイート数の期待値とした．(1) ジオタグが付いているツイート数の割合が極めて少ないため，ジオタグを用いた報酬の計算は困難であること，また (2) 発信地とツイートに出現する単語の関連が自明ではないため，特定トピックに関するツイートの収集手法 [9] と異なりツイート本文から報酬を直接計算できないこと，以上の 2 つの理由により，上田らはツイートのジオタグを考慮せずに，ツイートの発信地を推定することで報酬を算出した．ユーザ u をタイムウィンドウ $\acute{t}$ までにフォローした回数を $f_{u,\acute{t}}$，タイムウィンドウ $\acute{t}$ においてフォローユーザ $u_{\acute{t}}$ が発信したツイート集合を $TW_{u,\acute{t}}$ として，ユーザ $u_{\acute{t}}$ の評価は以下の 3 ステップで行う．

（1） ツイート $tw \in TW_{u,\acute{t}}$ が収集対象地域から発信された確率 $p(tw)$ を発信地推定器と CMN（後述）によって算出する．

（2） 確率 $p(tw)$ を集合 $TW_{u,\acute{t}}$ について合計することでタイムウィンドウ $\acute{t}$ 内のユーザ $u_{\acute{t}}$ による収集対象地域からの発信数の期待値 $e_{u,\acute{t}}$ を計算する．

$$e_{u,\acute{t}} = \sum_{tw \in TW_{u,\acute{t}}} p(tw) \tag{1}$$

（3） 期待値 $e_{u,\acute{t}}$ を，ユーザ $u_{\acute{t}}$ をフォローした過去のタイムウィンドウにおけるそれと合わせて平均値を計算する．その平均値をタイムウィンドウ $\acute{t}+1$ におけるユーザ $u_{\acute{t}}$ の報酬 $Q_{u,\acute{t}+1}$ とする．

$$Q_{u,\acute{t}+1} = \begin{cases} 0 & (f_{u,\acute{t}} = 0) \\ \frac{1}{f_{u,\acute{t}}} \sum_{t=1}^{\acute{t}} e_{u,\acute{t}} & (otherwise) \end{cases} \tag{2}$$

ツイートが収集対象地域から発信された確率を算出するために，分類器を用いる．学習データにはジオタグ付きツイートのみを用い，形態素の出現頻度からなる bag of words を 1 つのジオタグ付きツイートの特徴量ベクトルとする．クラスは発信地が対象地域であるか否かの二通りとする．この分類器の実運用段階では，収集対象地域から発信されるツイート数はその地域以外から発信されるツイート数に比べて著しく少ないため，class mass normalization（CMN）[20] を用いて分類器が出力する確率を補正する．CMN は分類器が出力する確率を実運用段階でのクラス数を考慮して補正する．例えば，クラス A，B の 2 クラスがありそれぞれの実際のクラス数の割合を r，$1-r$ とする．直感的には，ある分類器を用いてあるデータセットを分類した時に，クラス A，B それぞれの分類数 n_A，n_B の比率が $n_A : n_B = r : 1-r$ となるように CMN は分類器の出力確率を補正する．

5. 提案手法

本研究は第 4. 章で述べたバンディットアルゴリズムを用いた特定地域から発信されたツイートの収集手法をベースとし，新規ユーザの探索においてユーザ間のフォロー関係を Side information として用いることで，特定地域から発信されたツイート

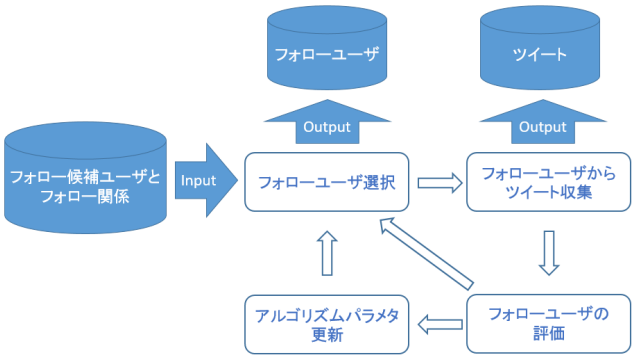


図 1: 提案手法の全体像

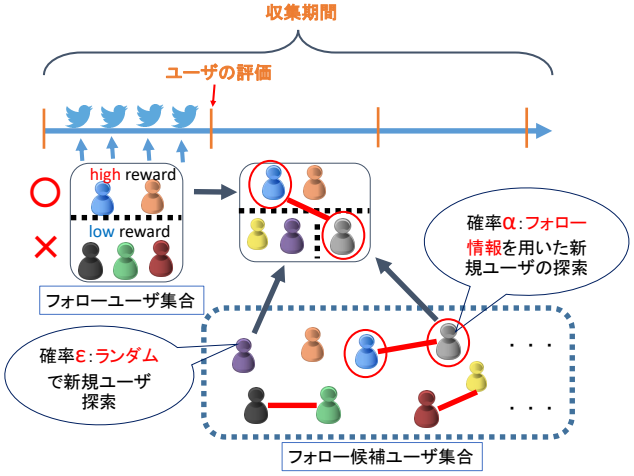


図 2: 提案手法 ϵ-α-greedy アルゴリズムの図解

の収集を効率化する．ユーザ間のフォロー情報は，フォロー候補ユーザを示すノードの集合 V_U，フォロー元ユーザ (follower) からフォロー先ユーザ (followee) へのフォローを示すエッジの集合 E_{follow} から構成される有向グラフ $G_{follow,t}(V_U, E_follow)$ とする．タイムウィンドウ t における収集の流れは以下の 4 つのステップである．なお初回タイムウィンドウのフォローユーザ集合はランダムで決定される．

（1） タイムウィンドウ t の開始時 : フォローユーザ選択（第 5.1 節）

以下を K（フォロー人数）回繰り返す

- 確率 ϵ : ランダムに新規ユーザをフォロー
- 確率 α : フォロー情報を用いて新規ユーザをフォロー
- 確率 $(1-\epsilon-\alpha)$: 評価値が最大のユーザをフォロー

（2） タイムウィンドウ t の間 : フォローユーザのツイートを収集

（3） タイムウィンドウ t の終了時 : フォローユーザの評価

（4） タイムウィンドウ $t+1$ へシフトしステップ 1. へ戻る（第 5.2 節）

I パラメタ ϵ,α はそのまま (ϵ-α-greedy-static 手法)

II パラメタ ϵ,α を更新 (ϵ-α-greedy-dynamic 手法)

ハイパーパラメタ Δ の値幅で ϵ を減少させ，α を増加させる．

- $\epsilon \leftarrow \epsilon - \Delta$
- $\alpha \leftarrow \alpha + \Delta$

5.1 フォローユーザ選択

本節では「フォローユーザ選択」について詳細に説明する．提案手法は新規ユーザの探索において，報酬と無関係にランダムで探索する方法と，直前タイムウィンドウにフォローしたローカルユーザのフォロー先にあたるユーザを重点的に探索する方法を組み合わせる．先行研究 [4] で使用された ϵ-greedy 手法へユーザのフォロー情報を組み込んだ提案手法を ϵ-α-greedy ツイート収集手法（以降，ϵ-α-greedy 手法）と呼ぶこととする．ϵ-α-greedy 手法のフォローユーザひとりの選択方法については，確率 ϵ で報酬とは無関係にランダムで探索フォローを行い，確率 α で直前のタイムウィンドウにおいてフォローしたユーザのフォロー先にあたるユーザをフォローし，確率 1-ϵ-α で報酬が最大のユーザをフォローする．各タイムウィンドウで上記 1 フォローユーザの選択を K 回繰り返すことでフォローユーザ集合を決定する．

ここで確率 α の場合に対応するフォロー情報を用いた 1 ユーザの選択方法は以下の 2 ステップとなる．ステップ 1 では直前のタイムウィンドウのフォローユーザ集合から報酬の高さに応じて確率的に 1 ユーザを選択し，ステップ 2 ではステップ 1 で選択したユーザのフォロー先にあたるユーザをランダムで選択しフォローする．

（1） $g_{u,t}$ をあるユーザ u のあるタイムウィンドウ t における報酬とする．

$\frac{g_{u,t}}{\sum_u g_{u,t}}$ の確率でフォロー元ユーザを選択．

（2） 上記ステップ 1 で選択されたフォロー元ユーザがフォローしているユーザの中からランダムで 1 ユーザを選択

5.2 パラメタ更新の有無

パラメタ ϵ, α の値をすべてのタイムウィンドウにおいて固定するものを ϵ-α-greedy-static 手法とする．この時 ϵ, α の固定値はハイパーパラメタとなる．これに加えて，1 タイムウィンドウごとに Δ の値幅で ϵ を減少させ，α を増加させるものを ϵ-α-greedy-dynamic 手法とする．これは収集時間を経るごとに発見済みローカルユーザ数が増加していく状況へ対応した新規ユーザの探索手法である．なお，ϵ の値の下限を th で設定する．ϵ の値が下限 th まで減少して以降は ϵ, α の値は固定となる．まとめると ϵ-α-greedy-dynamic 手法のハイパーパラメタは ϵ, α の初期値 ϵ_0, α_0 と値幅 Δ, ϵ の下限 th である．なお α の上限は $\epsilon_0 + \alpha_0 - th$ である．

6. 評価実験

評価実験では，提案手法がその他の手法に比べ，特定地域から発信されたツイートを効率よく収集できることを検証する．

6.1 実験データセット

実データを用いた実験により提案手法を既存手法と比較する．今回の実験では以下のデータセットを用いた．

6.1.1 フォロー候補ユーザ選定用データセット

フォロー候補ユーザ選定用データセットとして，Twitter データ提供サービスを利用し 2017 年の 1 月から 3 月に発信された

全ツイートのうち 10%をランダムで収集したツイートを用いる．このフォロー候補ユーザ選定用データセットに含まれるジオタグ付きツイートを多数発信したユーザ上位 2 万人をフォロー候補ユーザとして選定した．

6.1.2 ツイート収集手法のハイパーパラメタチューニング用データセット

Twitter Streaming API を利用して 2017 年の 5 月 26 日から 6 月 4 日にかけてフォロー候補ユーザのツイートを継続的に収集したデータをツイート収集手法のハイパーパラメタチューニング用データセットとする．このデータセット上で各ツイート収集手法を様々なハイパーパラメタ設定で動作させ，各ツイート収集手法について最も多くの目的ツイートを集めたパラメタを選定する．

6.1.3 本実験用データセット

Twitter Streaming API を利用して 2017 年の 6 月 16 日から 7 月 31 日にかけてフォロー候補ユーザのツイートを継続的に収集したデータを今回の本実験データセットとする．上記「ツイート収集手法のハイパーパラメタチューニング用データセット」で選定されたハイパーパラメタを設定値として各ツイート収集手法を動作させる．実際に得られた実験データはフォロー候補ユーザ 18466 人分のツイートである．フォローのネットワークはフォロー候補 18466 ユーザをノードとしたフォロー元からフォロー先への有向グラフである．リンクの出次数が 1 以上のノード数は 13768，フォローリンクの総数は 122,561 である．

6.1.4 発信地推定用学習データセット

発信地推定器の学習データとして，各対象地域についてその対象地域内とそれ以外の日本国内から発信されたジオタグ付きツイートを上記データセットとは別にそれぞれ 1 万件ずつ収集した．

6.2 実験設定

本実験では，収集する地域を 4 地域設定し，それぞれの地域について各手法でツイートを収集し，その有効性を比較する．ハイパーパラメタチューニング実験と本実験において，各手法・各ハイパーパラメタの組合せ・各地域の収集実験をそれぞれ 50 回ずつ行い，各評価指標値の平均値をとる．

対象地域はつくば市，東京 23 区，京都市，横浜市とした．タイムウィンドウ幅は先行研究 [4] と同様に 4 時間とした．ハイパーパラメタチューニング用，本実験用のデータセットのタイムウィンドウ数はそれぞれ 52，267 である．各タイムウィンドウにおいてフォローするユーザ数 K は，1,000，100 の 2 通りとした．フォローユーザ数の設定値を 2 通りとした理由は目的ツイート収集の難易度を変化させるためである．

発信地推定のための分類器にはナイーブベイズ分類器を用い，ツイートの特徴量 bag of words ベクトルには名詞のみを用いる．ツイートから名詞を抽出する際の形態素解析には MeCab を用いた．

6.3 評価指標

収集したツイートの中にはジオタグが付与されていないものが多いため，収集したジオタグ付きツイートを用いて特定地域から発信されたツイートの収集数を推定する．先行研究 [4] で用いられた評価指標である．その評価指標の計算手順を以下に示す．

（1） タイムウィンドウ t において 1 フォローユーザ u_t から収集した目的ツイート数を推定する．

（2） タイムウィンドウ t までに全フォローユーザから収集した目的ツイート推定数を合計したものを，タイムウィンドウ t における特定地域から発信されたツイートの収集数 $n_t^{estimated}$ とする．

1 フォローユーザ $u_{\acute{t}}$ から収集した目的ツイート数の推定手順は以下の通りである．

（1） フォローしたユーザを場合分けする．

I $u_{\acute{t}} \in U_{\acute{t},A}$：タイムウィンドウ $\acute{t}$ においてジオタグ付きツイート $p \in P_{public}$ を 1 件以上発信した．

II $u_{\acute{t}} \in U_{\acute{t},B}$：タイムウィンドウ $\acute{t}$ においてジオタグ付きツイート $p \in P_{public}$ を発信していない．

（2） 場合分けに応じて 1 ユーザごとに収集数を推定する．

I $u_{\acute{t}} \in U_{\acute{t},A}$：タイムウィンドウ $\acute{t}$ においてユーザが 1. 発信したツイート数と 2. 対象地域から発信したジオタグ付きツイートの割合から推定する．

$$n_{\acute{u},\acute{t}} = \frac{|\{tw_i \in TW_{public} | u=\acute{u}, t=\acute{t}, l=\tilde{l}\}|}{|\{tw_i \in TW_{public} | u=\acute{u}, t=\acute{t}\}|} |\{tw_i \in TW | u=\acute{u}, t=\acute{t}\}| \quad (3)$$

II $u_{\acute{t}} \in U_{\acute{t},B}$：タイムウィンドウ $\acute{t}$ における 1. $U_{\acute{t},A}$ の総ツイート数に対する $n_{u,\acute{t}}$ の合計の割合と，2. $u_{\acute{t}}$ のツイート数から推定する．

$$n_{\acute{u},\acute{t}} = \frac{\sum_{u \in U_{\acute{t},A}} n_{u,\acute{t}}}{|\{tw_i \in TW | u \in U_{\acute{t},A}, t=\acute{t}\}|} |\{tw_i \in TW | u=\acute{u}, t=\acute{t}\}| \quad (4)$$

6.4 比較手法

今回の実験ではベースラインとなる既存手法と提案手法とを比較評価した．ベースラインとして用いた手法は以下の 4 つの手法である．

- Random：ランダムにフォローユーザを選びツイートを収集する．
- StatisticsNTW：最初の N タイムウィンドウまではランダムに選出した NK ユーザを順番に一度ずつフォローし，$N+1$ タイムウィンドウ以降は N タイムウィンドウまでに得られた報酬上位 K ユーザをフォローし続ける．
- Number：ツイート数の多いユーザをフォローすることでツイートを収集する．発信ツイート数を報酬とする ϵ-greedy アルゴリズムを用いてフォローユーザを選出する．
- ϵ-greedy：特定地域からの発信ツイート数を報酬とする ϵ-greedy アルゴリズムを用いてフォローユーザを選出する．

6.5 収集手法のハイパーパラメタ選定

今回の既存手法と提案手法のハイパーパラメタ値の候補を表 2 にまとめる．なお $r_{explore} = \epsilon + \alpha$ とするハイパーパラメタ $r_{explore}$ を導入している．本実験とは別のデータセットを用いて，ハイパーパラメタ値の候補から $K = 1000, 100$ のそれぞれについて収集手法の最適なハイパーパラメタを本実験のために選定した．

表 2: 比較手法とハイパーパラメタ値の候補

比較手法	ハイパーパラメタ値の候補
ϵ-α-greedy-static	$r_{explore} \in \{0.3,\ 0.5,\ 0.7\}$, $\epsilon \in \{0,\ r_{explore} \times 1/4,\ r_{explore} \times 1/2,\ r_{explore} \times 3/4,\ r_{explore}\}$, $\alpha = r_{explore} - \epsilon$
ϵ-α-greedy-dynamic	$r_{explore} \in \{0.3,\ 0.5,\ 0.7\}$, $\epsilon_0 \in \{r_{explore} \times 1/2,\ r_{explore} \times 3/4,\ r_{explore}\}$, $\alpha_0 = r_{explore} - \epsilon_0$, $\Delta \in \{0.1,\ 0.05,\ 0.01\}$, $th \in \{0.1,\ 0.05,\ 0\}$
ϵ-greedy Number	$\epsilon \in \{0.3,\ 0.5,\ 0.7\}$
Random Statistics6TW Statistics18TW	ハイパーパラメタ無し

6.6 本実験の結果と考察

6.6.1 特定地域から発信されたツイートの推定収集数

各収集手法による特定地域から発信されたツイートの推定収集数 $n_t^{estimated}$ の推移の平均値を $K = 1000,\ 100$ の場合についてそれぞれ図 3，4 に示す．

$K = 1000$ の場合，図 3 が示す通り，全地域で提案手法による最終的な推定収集数が既存手法によるそれを上回っている．特に序盤の 50 タイムウィンドウでの提案手法と既存手法による推定収集数の差が広がっている．これはフォロー情報を用いて新規フォローユーザの探索範囲を限定することで，目的のツイートを発信するユーザをより早く発見できることを示している．

しかしながら，中盤のタイムウィンドウ以降京都市は他の 3 地域と異なり，提案手法と既存手法の一つ ϵ-greedy 手法による推定収集数の差が広がっていないことが見てとれる．京都市の場合，発信するユーザの内観光客が占める割合が大きいと考えられる．一時的な滞在者によるツイート数が多い地域においてユーザの交友関係を示すグラフを用いた提案手法の収集性能については今後さらなる検討が必要である．

$K = 100$ の場合，$K = 1000$ の場合と比べ提案手法と既存手法の収集性能差が大きくなった．$K = 100$ の場合，特定地域から発信するユーザを発見することがより困難になる．厳しい収集条件のもとでは，フォロー情報を用いた提案手法が特に有効であることを示している．

しかしながら，次の第 6.6.2 節で見るように，この厳しい収集条件を一定水準の人口に満たないつくば市のような地域に適応した時，提案手法による目的ツイート収集が失敗する可能性がある．

6.6.2 推定収集数とその標準偏差

特定地域から発信されたツイートの最終的な推定収集数 $n_T^{estimated}$ の 50 回平均とその標準偏差を収集手法間で比較すると，$K = 1000,\ 100$ の場合についてそれぞれ図 5，6 の通りである．

$K = 1000$ の場合，提案手法による推定収集数の標準偏差が既存手法と比べ小さいことが分かる．ランダムな新規フォローユーザの探索と比べ，探索フォロー情報を用いた探索は，目的のツイートを発信するユーザをより確実に発見できることを示している．

$K = 100$ の場合，つくば市において提案手法による推定収集数の標準偏差が既存手法と比べ大きいことが分かる．一度にフォロー可能なユーザ数が少ないという収集条件に加え，収集対象地域の人口が一定水準に満たない場合，フォロー情報を用いて新規フォローユーザの探索範囲を限定することが，目的ツイート収集失敗の要因になると考えられる．厳しい収集条件における，収集失敗の可能性を低減することが今後の検討項目となる．

7. ま と め

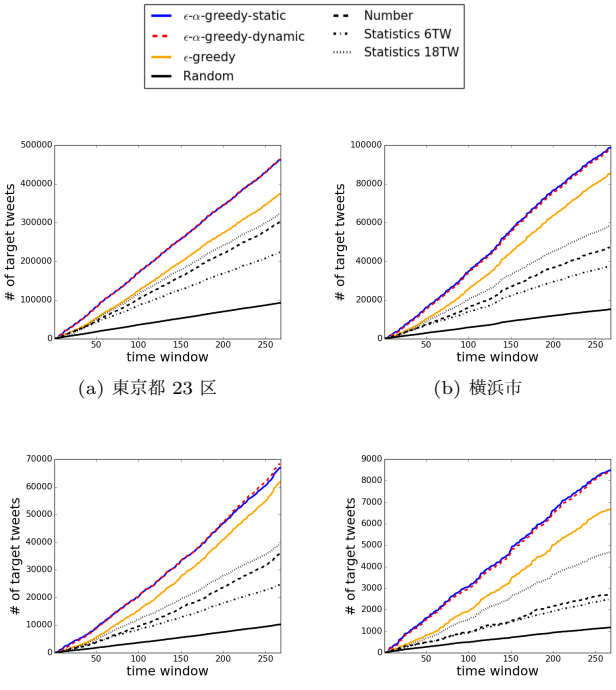
本研究は Twitter におけるユーザ間のフォロー関係の情報を用いて，特定地域より発信されたツイートを効率的に収集する手法を提案した．ユーザ間のフォロー情報を新規ユーザの探索に用いることで，特定地域から発信されたより多数のツイートを収集できることを，実データを用いて示した．新規ユーザの探索へユーザ間のフォロー情報を用いることで目的のツイートをより多く発信しそうなユーザをより短期間で発見できツイート収集を効率化できた．しかし収集対象地域の人口や一度にフォロー可能なユーザ数が限定される厳しい収集環境ではフォロー情報を用いた目的ツイートの収集が失敗する場合があることを確認した．今後，厳しい収集環境における収集失敗を詳細に分析するために様々な人口規模を持つ都市での追加収集実験を行い，収集失敗の回避方法を検討することが必要である．またツイートの宛先，リツイートといった情報の導入や発信地推定性能の改善手法の検討も必要である．

8. 謝　　辞

本研究の一部は，文科省/理研「実社会ビッグデータ利活用のためのデータ統合・解析技術の研究開発」による．

文　　献

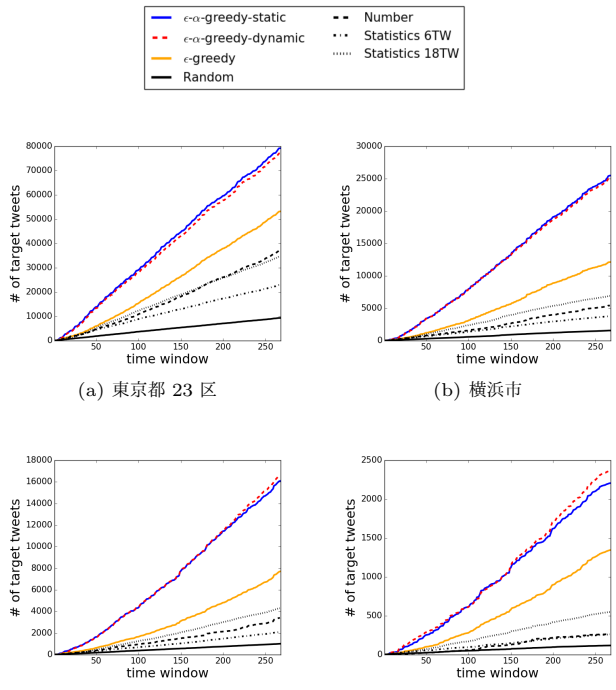
[1] Jianshu Weng, Yuxia Yao, Erwin Leonardi, and Francis Lee.



(a) 東京都 23 区　(b) 横浜市

(c) 京都市　(d) つくば市

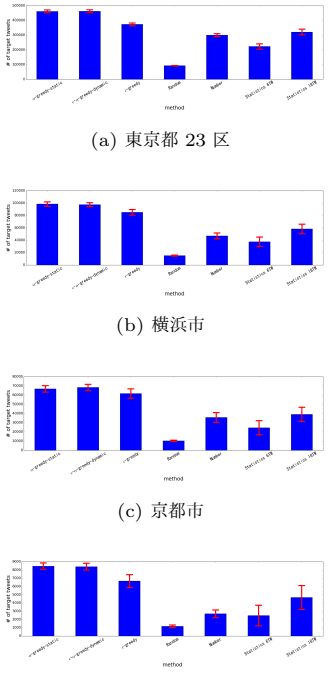
図 3: 目的ツイート推定収集数 $n_t^{estimated}$ の推移（$K = 1000$, 267 タイムウィンドウ）．x 軸はタイムウィンドウを表し，y 軸は各タイムウィンドウまでの対象地域から発信されたツイートの累計収集数を表す．全地域で提案手法が比較手法を最終的なツイート収集数で上回っていることが分かる．



(a) 東京都 23 区　(b) 横浜市

(c) 京都市　(d) つくば市

図 4: 目的ツイート推定収集数 $n_t^{estimated}$ の推移（$K = 100$, 267 タイムウィンドウ）．x 軸はタイムウィンドウを表し，y 軸は各タイムウィンドウまでの対象地域から発信されたツイートの累計収集数を表す．



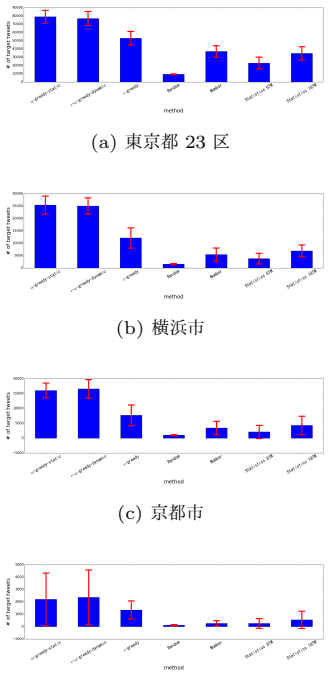
(a) 東京都 23 区

(b) 横浜市

(c) 京都市

(d) つくば市

図 5: 最終的な目的ツイート推定収集数 $n_T^{estimated}$ の比較（$K = 1000$，267 タイムウィンドウ）．x 軸は各収集手法を，y 軸は対象地域から発信されたツイートの最終的な収集数を表す．



(a) 東京都 23 区

(b) 横浜市

(c) 京都市

(d) つくば市

図 6: 最終的な目的ツイート推定収集数 $n_T^{estimated}$ の比較（$K = 100$，267 タイムウィンドウ）．x 軸は各収集手法を，y 軸は対象地域から発信されたツイートの最終的な収集数を表す．提案手法の収集数の標準偏差が大きいことが分かる．

Event detection in twitter. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, 2011.

[2] Stuart E Middleton, Lee Middleton, and Stefano Modafferi. Real-time crisis mapping of natural disasters using social media. *IEEE Intelligent Systems*, Vol. 29, No. 2, pp. 9–17, 2014.

[3] Kalev Leetaru, Shaowen Wang, Guofeng Cao, Anand Padmanabhan, and Eric Shook. Mapping the global twitter heartbeat: The geography of twitter. *First Monday*, Vol. 18, No. 5, 2013.

[4] Saki Ueda, Yuto Yamaguchi, and Hiroyuki Kitagawa. Collecting non-geotagged local tweets via bandit algorithms. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, CIKM 2017, Singapore, November 06 - 10, 2017*, pp. 2331–2334, 2017.

[5] Herbert Robbins. Some aspects of the sequential design of experiments. *Bull. Amer. Math. Soc.*, Vol. 58, No. 5, pp. 527–535, 09 1952.

[6] Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Mach. Learn.*, Vol. 47, No. 2-3, pp. 235–256, May 2002.

[7] Christopher John Cornish Hellaby" "Watkins. *"Learning from Delayed Rewards"*. PhD thesis, "King's College", "Cambridge, UK", "May" "1989".

[8] Shipra Agrawal and Navin Goyal. Analysis of thompson sampling for the multi-armed bandit problem. In *COLT*, pp. 39–1, 2012.

[9] Thibault Gisselbrecht, Ludovic Denoyer, Patrick Gallinari, and Sylvain Lamprier. Whichstreams: A dynamic approach for focused data capture from large social media. In *Proceedings of the Ninth International Conference on Web and Social Media, ICWSM 2015, University of Oxford, Oxford, UK, May 26-29, 2015*, pp. 130–139, 2015.

[10] Chih-Chun Wang, S. R. Kulkarni, and H. V. Poor. Bandit problems with side observations. *IEEE Transactions on Automatic Control*, Vol. 50, No. 3, pp. 338–355, March 2005.

[11] Omid Madani and Dennis DeCoste. Contextual recommender problems [extended abstract]. In *Proceedings of the 1st International Workshop on Utility-based Data Mining*, UBDM '05, pp. 86–89, New York, NY, USA, 2005. ACM.

[12] Lihong Li, Wei Chu, John Langford, and Robert E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pp. 661–670, New York, NY, USA, 2010. ACM.

[13] Swapna Buccapatnam, Atilla Eryilmaz, and Ness B. Shroff. Stochastic bandits with side observations on networks. In *The 2014 ACM International Conference on Measurement and Modeling of Computer Systems*, SIGMETRICS '14, pp. 289–300, New York, NY, USA, 2014. ACM.

[14] Yohei Ikawa, Miki Enoki, and Michiaki Tatsubori. Location inference using microblog messages. In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12 Companion, pp. 687–690, New York, NY, USA, 2012. ACM.

[15] Clodoveu A. Davis Jr., Gisele L. Pappa, Diogo Renn Rocha de Oliveira, and Filipe de L. Arcanjo. Inferring the location of twitter messages based on user relationships. *Transactions in GIS*, Vol. 15, No. 6, pp. 735–751, 2011.

[16] Axel Schulz, Aristotelis Hadjakos, Heiko Paulheim, Johannes Nachtwey, and Max Mhlhuser. A multi-indicator approach for geolocalization of tweets. In *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*, 2013.

[17] Zhiyuan Cheng, James Caverlee, and Kyumin Lee. You are where you tweet: A content-based approach to geo-locating twitter users. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, CIKM '10, pp. 759–768, New York, NY, USA, 2010. ACM.

[18] Adam Sadilek, Henry Kautz, and Jeffrey P. Bigham. Finding your friends and following them to where you are. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*, WSDM '12, pp. 723–732, New York, NY, USA, 2012. ACM.

[19] Rui Li, Shengjie Wang, Hongbo Deng, Rui Wang, and Kevin Chen-Chuan Chang. Towards social user profiling: Unified and discriminative influence model for inferring home locations. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '12, pp. 1023–1031, New York, NY, USA, 2012. ACM.

[20] Xiaojin Zhu, Zoubin Ghahramani, John Lafferty, et al. Semi-supervised learning using gaussian fields and harmonic functions. In *ICML*, Vol. 3, pp. 912–919, 2003.