

バンディットアルゴリズムとメンション関係を利用した 特定トピックに関する特定の地域のツイートの収集

大森雄基[†] 北川博之[‡] 天笠俊之[‡]

[†]筑波大学 情報科学類 [‡]筑波大学 計算科学研究センター

1 はじめに

情報機器の一般化に伴い SNS の利用は一般化し、これを利用して地域に密着したマーケティングや商圏分析等に使用することが一般化している。このような背景のもと、特定の地域から特定のトピックに沿った投稿を効率よく集める需要が存在する。しかし、SNS 上では、様々な人々が様々な話題を投稿し、しかも情報は極めて膨大で断片的であるため、人間がそれをまとめるのは容易ではない。また、Twitter を含む SNS の投稿情報は自由に得られるものではなく、計算機の扱えるデータ量・計算力の制約も存在する。また、Twitter のプロフィールには居住地欄があるが、正しくない住所や無意味な言葉が入っている場合も多く [3]、ツイート全体に占める Geotag 付きツイートも 0.7% しかない [4]。このため、ユーザの活動地域を推定しつつ、投稿を効率よく収集するためのアルゴリズムが求められる。

本研究は、Twitter において特定の地域で活動する人から発信された特定のトピックに関するツイートを効率よく集めるという問題に対し、バンディットアルゴリズムに基づいて適切なユーザを選んでツイートを収集する機構を提案する。ツイート本文だけではなくプロフィールも利用したユーザの活動地域推定とトピック文章類似度を用いた第一報酬、メンション関係と言及元ユーザの第一報酬を利用した第二報酬、活動地域推定のみを利用した第三報酬の、三つの報酬を用いてユーザを選択し、ツイートを収集する。

2 関連研究

バンディットアルゴリズムとは、複数の選択肢に事前に知ることのできない報酬が定められ、選択肢を選ぶことで報酬を受け取る行為が一定の回数しかできな

い状況下において、累積報酬を最大化することを目的とするアルゴリズムである。この手法は活用と探索から成り、活用はその一回の選択で得られる報酬を最大化する行動であり、探索は選択肢の報酬の分布を知るための行動である。 ϵ -greedy 法は、探索と活用の方策の一種であり、 $1 - \epsilon$ の確率で活用を、 ϵ の確率で探索としてランダムに選択する。

先行研究においては特定地域から発信されたツイートの収集という問題に対してバンディットアルゴリズムを適用している [1][2]。時間をタイムウィンドウごとに区切り、それぞれのタイムウィンドウで個々のツイートに対する発信地推定に基づいて各ユーザの推定報酬を計算し、これに基づいて収集対象のユーザを決定する。本研究では、発信地推定とは異なる問題である活動地域推定を扱う。これは、個々のツイートではなく、ユーザに対してその活動している地域を推定するものである。

3 提案手法

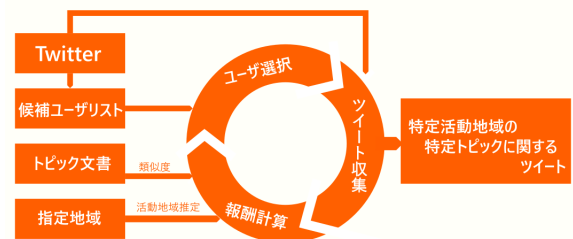


図 1: 本研究の概要：タイムウィンドウのサイクル

本研究では、以下の手順でツイートを収集する。

1. K 人のユーザを以下の方法で選択し、フォロー。
(各報酬の詳細は後述)
 - a 確率 ϵ_1 で第一報酬 $Q_1(u, \tau_f)$ の高いユーザを選択
活動地域推定・トピック文書類似度より計算
 - b 確率 ϵ_2 で第二報酬 $Q_2(u, \tau_f)$ の高いユーザを選択
メンション数と言及元ユーザの第一報酬
 - c 確率 ϵ_2 で第三報酬 $Q_3(u, \tau_f)$ の高いユーザを選択
活動地域推定より計算

Method for Collecting Topic-related Local Tweets using Bandit Algorithms and Mention Relationships
Yuuki OHMORI[†], Hiroyuki KITAGAWA[†] and Toshiyuki AMAGASA[‡]

[†]College of Information Science, University of Tsukuba
1-1-1 Tennodai, Tsukuba 305-8573, Japan

[‡]Center for Computational Sciences, University of Tsukuba
1-1-1 Tennodai, Tsukuba 305-8577, Japan

[†]omori.yuki.sm@alumni.tsukuba.ac.jp [‡]{kitagawa, amagasa}@cs.tsukuba.ac.jp

- d 確率 $1 - (\epsilon_1 + \epsilon_2 + \epsilon_3)$ でランダムに取得
2. ユーザからタイムウィンドウ中のツイートを取得.
3. ユーザごとに推定報酬やスコア・得点を算出.

各ユーザ u について以下に述べる方法でタイムウィンドウ τ_f ごとの推定報酬 $q_1(u, \tau_f), q_2(u, \tau_f), q_3(u, \tau_f)$ を計算し、それらのタイムウィンドウ τ_f までの平均値として、累積推定報酬 $Q_1(u, \tau_f), Q_2(u, \tau_f), Q_3(u, \tau_f)$ を算定する.

第一報酬 $q_1(u, \tau_f)$ はタイムウィンドウ τ_f に取得したユーザ u のツイートの集合 $X_u^{\tau_f}$ に基づいて算出したトピック文書類似度 Sim と、活動地域推定から算出した第三報酬 $q_3(u, \tau_f)$ を用いて算出する.

$$q_1(u, \tau_f) = q_3(u, \tau_f) \times \sum_{x \in X_u^{\tau_f}} \text{Sim}(\text{vec}_t(x))$$

第二報酬は、同じ話題を話しているか、居住地に近いならば相手ユーザにメンションされる可能性の高いという考えのもと、タイムウィンドウ中にユーザ u が u' にメンションされた回数 $\text{Mention}(u' \rightarrow u, \tau_f)$ 、メンション探索元ユーザのタイムウィンドウ内における第一報酬 $q_1(u', \tau_f)$ を用いて次のように計算する.

$$q_2(u, \tau_f) = \sum_{u' \in U^{\tau_f}} \text{Mention}(u' \rightarrow u, \tau_f) \times q_1(u', \tau_f)$$

第三報酬は、タイムウィンドウ τ_f までに取得したユーザ u のツイートの集合 $X_u^{\tau_f} = \bigcup_{i=1}^f X_u^{t_i}$ と、ユーザプロフィールの `description` 欄と `location` 欄から、分類器を用いて該当地域で活動している確率を算出し、これを用いる.

$$q_3(u, \tau_f) = \text{Cls}_a^{\text{location}} \left(\begin{bmatrix} \frac{1}{|X_u^{\tau_f}|} \sum_{x \in X_u^{\tau_f}} \text{Cls}_t^{\text{location}}(\text{vec}_t(x)) \\ \text{vec}_d(\text{profDesc}_u) \\ \text{vec}_l(\text{profLoc}_u) \end{bmatrix} \right)$$

4 実験・考察

2016 年 12 月から 2017 年 4 月にツイートされたツイートとユーザプロフィールを用い、geotag 付ツイートの多いユーザ 3 万を選び、その中から 2 万ユーザをランダムに選出し、ツイート収集機構全体と各種活動地域推定分類器の試験用データとした. また、残りの 1 万ユーザから、対象地域と非対象地域のユーザを同数ずつ抽出してユーザ活動地域分類器 $\text{Cls}_a^{\text{location}}$ の学習用データとした. 前記 1 万ユーザのツイートを含め、試験用データ以外のツイートから対象地域と非対象地域を同数ずつ選出してツイート活動地域分類器 $\text{Cls}_t^{\text{location}}$ とした. 分類器は scikit-learn の `RandomForestClassifier` を使用した. 文字列は Mecab と gensim を用いて単語の頻度ベクトルとした. 評価基準は、(ユーザの該当地域の geotag 付きツイート / ユーザの geotag 付きツ

weet) $\times$ (収集したツイートの文書類似度の合計) の積算である. 比較手法は、それぞれの報酬の ϵ を変更するだけでなく、D 日で収集対象ユーザを固定したもの ($D = 1, 2, 3$), トピック文書類似度のみを報酬としたもの、ツイート数のみを報酬にしたもの、また完全にランダムのもを用意した. 比較手法を含め、結果は 20 回づつ試行した平均をとった.

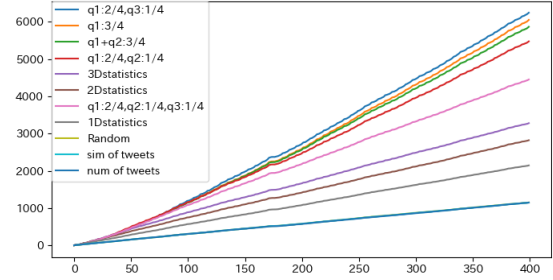


図 2: 結果の一部 | 地域: 京都市, トピック: 築地, $K = 1000$ | 凡例配列は最終評価値順

以上の結果から、第一報酬と第三報酬によるツイート収集はどの比較手法よりも特定トピックに関する特定の地域のツイートの収集できることが分かった. また、メンションを利用した場合、この設計では性能を引き下げてしまうことが判明した.

5 まとめ

本研究では、活動地域推定とトピック文書類似度を用いてユーザを選択し、ツイートを収集する手法を提案した. 今後の課題は、メンションによる引き下げ効果の原因究明と、新たな第二報酬の設計である.

参考文献

- [1] Carina Miwa Yoshimura and Hiroyuki Kitagawa, TLV-Bandit: Bandit Method for Collecting Topic-related Local Tweets, Proc. MIPR2021
- [2] 上田紗希, 山口祐人, 北川博之, 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, DBSJ Japanese Journal, Vol.16-J, Article No.09
- [3] Brent Hecht, et al., Tweets from Justin Bieber's Heart: The Dynamics of the "Location" Field in User Profiles, Proc. Conf. on Human Factors in Computing Systems, 2011, pp. 237-246.
- [4] Mark Graham, Scott A Hale, and Devin Gaffney. 2014. Where in the world are you? geolocation and language identification in twitter. The Professional Geographer, 66(4):568-578.