

Twitter ユーザの活動するローカル地域の発見 Local Area Discovery of Twitter Users

田原琢士[†] 馬強[†]
Takuji Tahara Qiang Ma

京都大学大学院情報学研究科[†]
Graduate School of Informatics
Kyoto University

1. はじめに

代表的なマイクロブログである Twitter¹は約 2.7 億人²の月間アクティブユーザを持ち、ユーザの投稿のリアルタイム性とそのオープン性から多くの研究者に情報源として注目されている。Twitter に関する研究にはマーケティングや情報推薦、広告に役立つユーザの属性を推定する研究が数多く存在し、特にユーザの生活地域を推定する研究は、Twitter に備わっている位置情報付与機能の利用率が僅か (2012 年の時点で 0.77%³) であるために非常に盛んである。

しかし、ユーザの生活地域を推定する既存研究 ([1][2][3]など) の多くは、ユーザを国や州、県、市といった行政区に分類するのが主であり、推定した生活地域がユーザの実際に活動しているエリアにマッチしているとは限らない。

そこで本稿では、ユーザの生活地域推定をより小さな粒度で行う為の領域、ローカル地域を発見する手法を提案する。具体的には、まずジオタグの付いた Tweet を頻繁に投稿しているユーザ個人の活動領域を推定し、それらのユーザの活動領域の地理的な重なりを考慮して、生活圏のようなユーザが活動する地域を生成する。

2. 提案手法

本手法では、まずジオタグ付き Tweet を頻繁に投稿するユーザ各個人の活動している領域の推定を行う (2.1 節)。次に、求めた個人の活動領域を地理的な類似性から統合し、求めるローカル地域を生成する (2.2 節)。

2.1. ユーザ個人の活動領域の推定手法

ユーザ個人の活動領域は一つではないと考えられる。例えば、自宅と職場が離れているユーザ

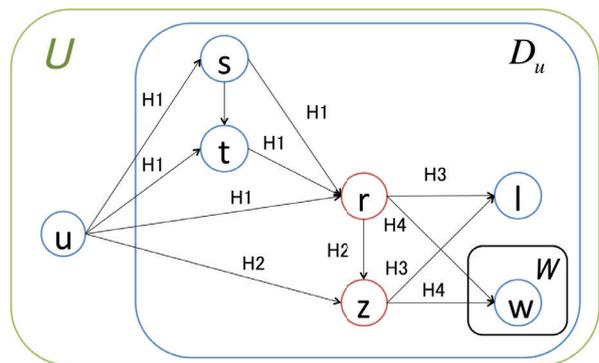


図 1: W^4 のグラフ (出典 [4])。

Tweet は $d = (u_d, s_d, t_d, l_d, w_d)$ で表され、 u は投稿したユーザ、 s は投稿されたのが週末か否か、 t は投稿された時間、 l_d は投稿された位置、 w は投稿に含まれる単語を表す。また、 r, z は隠れ変数である個人領域とトピックを示す。

は自宅と職場それぞれを中心とする 2 つの活動領域を持つ。この複数の活動領域を推定するために、本手法では Yuan ら [4] が提案した確率モデル W^4 を利用する。 W^4 は以下の四つの直感的仮説に基づいて Tweet を生成するモデルである。(図 1)

仮説 H1 : 各ユーザは行動の軸となるいくつかの個人領域を持っており、それらの領域に含まれる場所を訪れる傾向がある。また、ユーザがどの個人領域に居るかは、時間的要素 (例: 日中か夜か、平日か週末か) に影響される。

仮説 H2 : ユーザのトピックは、そのユーザのトピックの嗜好と現在居る個人領域に影響される。

仮説 H3 : ユーザを訪れる場所は、ユーザのトピック要件と、ユーザの現在居る個人領域に影響される。

仮説 H4 : ユーザの投稿内容 (単語) は現在居る個人領域とユーザのトピックに影響される。

W^4 はユーザごとの個人領域とトピックを隠れ変数としており、ジオタグ付き Tweet を用いて最尤推定を行うことでこれらを求めることができる。

¹ <https://twitter.com/>

² <https://biz.twitter.com/ja/whos-twitter>

³ <http://semicast.com>

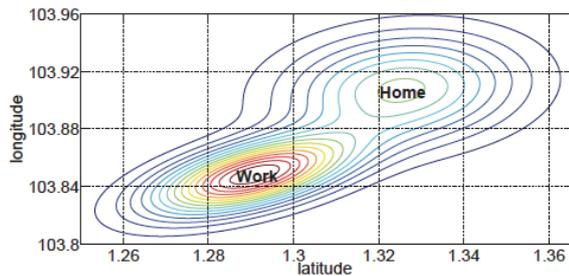


図 2: ユーザの個人領域の例 (出典:[4])

ユーザ u_i の j 番目の個人領域 $r_{u_i,j}$ は、平均 $\mu_{r_{u_i,j}}$ 、分散行列 $\Sigma_{r_{u_i,j}}$ の二次元正規分布 $f_{r_{u_i,j}}(x)$ で表され、あるユーザの複数の個人領域は図 2 に示したような等高線で可視可される。我々はこの個人領域をユーザ個人の活動領域として扱う。

2.2. ローカル地域の生成

直感的には、複数のユーザの活動領域が重なっていればそれらを統合して一つのローカル地域を生成する。すなわち 2.1 節で求めた各ユーザの活動領域を表す二次元正規分布 $f_r(x)$ を用いて二次元混合正規分布 $F(x)$ を (1) のように生成し、 $F(x)$ の値による等高線図の等高線が密になっている領域をローカル地域と見なす。

$$F(x) = \sum_r f_r(x) \quad (1)$$

図 3 に $F(x)$ の等高線図のイメージ図を示す。ただし、本稿では実際に呟かれた Tweet の座標を通る等高線のみを用いる。図 3 では高さの異なる 2 つの山が出来ており、直感的にはこれらがそれぞれローカル地域となる。しかし、単純に一定の高さ以上の領域をローカル地域とすると、高さが異なることから消滅してしまうローカル地域が出たり、複数のローカル地域が一つだと見なされたりする可能性がある。そこで、 $F(x)$ の各山 (M_k とする) ごとに、隣接する等高線との落差が急変している等高線の、一段低い等高線で区切ることによりローカル地域を生成する。まず各 M_k の頂点 c_k 、すなわち $F(x)$ の極大点を求め、2.1 節で求めた各ユーザの活動領域 r を、その中心 (平均) μ_r が最も近い c_k を持つ M_k に割り当てる。次に各 M_k に属する r で実際に呟かれた Tweet の座標の集合 $L_k = (l_{k,1}, \dots, l_{k,n_k})$ を用いて、以下のようにして $l_{k,i} \in L_k$ における落差の変化量 $dh(l_{k,i})$ を計算する。

$$dh(l_{k,i}) = \frac{F(l_{k,i}) - F(l_{k,i}^-)}{F(l_{k,i}^+) - F(l_{k,i})} \quad (2)$$

ここで、 $l_{k,i}^+$ 、 $l_{k,i}^-$ は以下のような座標である。

$$l_{k,i}^+ = \operatorname{argmin}_{l \in L_k \wedge \{F(l_{k,i}) < F(l)\}} F(l) \quad (3)$$

$$l_{k,i}^- = \operatorname{argmax}_{l \in L_k \wedge \{F(l_{k,i}) > F(l)\}} F(l) \quad (4)$$

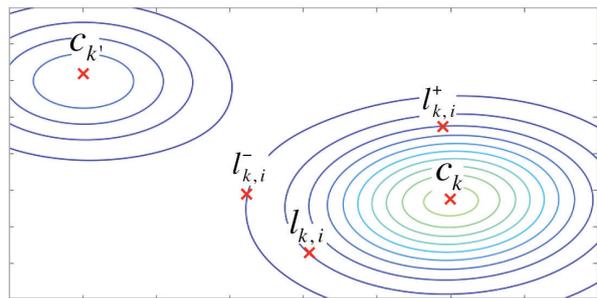


図 3: $F(x)$ の等高線図のイメージ図

各 $l_{k,i} \in L_k$ に対し、 $F(l_{k,i})$ の値が大きい順に $dh(l_{k,i})$ を計算していき、閾値 θ について $dh(l_{k,i}) < \theta$ を満たせば高さ $F(l_{k,i})$ で落差が急変しているの見なし、一段低い $F(l_{k,i}^-)$ の等高線をローカル地域の外周とする。この計算を各 M_k に対して行ってローカル地域を生成する。

3. まとめ

本稿では Twitter におけるユーザの生活地域推定をより小さな粒度で行う為のローカル地域を生成する手法を提案した。

今後の実験計画としては、Eisenstein ら [5] の提供している Geo-tagged Microblog Corpus を用いて実験を行い、地理学上の生活圏などと比較して妥当性を検討する。更に追加実験として、作成したローカル地域で呟かれたと推定した Tweet のテキストに我々の先行研究 [6] を適用し、生活地域推定の評価実験を行う予定である。

参考文献

- [1] Z. Cheng, J. Caverlee, K. Lee. You Are Where You Tweet: A Content-based Approach to Geo-locating Twitter Users. In *CIKM*, pp. 759-768 (2010).
- [2] 堂前友貴, 関洋平. 地域に偏りのあるトピックを用いた Twitter ユーザの生活に関わる地域推定. 情報処理学会・情報学基礎研究会報告, Vol. 2013, No. 8, pp. 1-6 (2013).
- [3] 西村駿人, 数原良彦, 鷲崎誠司. 地域特徴語選択を用いたマルチクラス分類による Twitter ユーザの居住地推定. 信学技報, Vol. 112, No. 367, pp. 23-27 (2012).
- [4] Q. Yuan, G. Cong, Z. Ma, A. Sun, N. Magnenat-Thalmann. Who, where, when and what: discover spatio-temporal topics for twitter users. In *KDD* pp. 605-613 (2013).
- [5] J. Eisenstein, B. O'Connor, N. A. Smith, E. P. Xing. A Latent Variable Model for Geographic Lexical Variation. In *EMNLP*, pp. 1277-1287 (2010).
- [6] T. Tahara, Q. Ma. Searching for Local Twitter Users by Extracting Regional Terms. In *DEXA*, pp. 89-96 (2014).