ARG WI2 No.13, 2018
WI2-2018-16

東京近郊を対象としたTwitterユーザの細粒度ロケーション予測に関する検討

松野 省吾, 水木 栄, 榊剛史

株式会社ホットリンク

shogo.matsuno@hottolink.co.jp

概要 SNS 広告需要の拡大を受け, より高精度なターゲティング手法が求められている. そこで, 筆者らはエリアマーケティングを想定した Twitter ユーザの細粒度ロケーション予測を試みた. 具体的には, ジオタグ付き Tweet などから得られる住所情報をラベルとしたラベル伝搬アルゴリズムを用いて街区ごとの訪問確率を予測し, 大字レベルの粒度で該当エリアを訪れる可能性のある Twitter ユーザの予測を目指す. 検証実験として, 提案手法に基づき街区名を入力すると当該地点を訪れる可能性の高いユーザのリストを出力するシステムを構築し, 1 都 4 県に属する市区町村, および大字を基準とした約 8000 街区に対し, 各街区をユーザが訪れる可能性の予測を試みた. その結果, 全体を平均して都道府県レベルでの予測精度は 73%, 市区町村レベルでは 42%, 大字レベルでは 25%となった. また, 東京都のみで予測したところ, 大字レベルで 31%の予測精度となった. これらの結果を踏まえ, 提案手法の有効性について考察する.

キーワード Twitter, ロケーション予測, エリアマーケティング, ターゲティング, ラベル伝搬

1 はじめに

インターネット広告媒体費は年々拡大しており, 効率的な広告配信を行うためのよりよいターゲティング手法が求められている. こうした高精度, 高粒度なターゲティングを行う方法のひとつとして, ジオターゲティングが挙げられる[1, 2]. ジオターゲティングとは, 位置情報からユーザの居場所を解析し, セグメント化することで地域に特化したエリアマーケティングを実施する手法である. 一般的に, ジオターゲティングに用いられるロケーション予測方法は, ユーザのスマートフォンなどから取得した位置情報履歴に基づく方法とWebサービスの利用履歴に基づく方法が挙げられる. 前者はジオフェンシングと呼ばれ, 任意の特定エリアをあらかじめ設定し, エリアに出入りがあった場合に広告を配信することができる. 一方で, 位置情報を取得するための物理センサの所持が前提であること, 現在までに実際に設定エリアに訪れたことのあるユーザしか対象とできないことなどの課題がある. 後者の方法はこうした制約はないものの, 精度が Web サービスを利用するユーザの登録情報に依存することや, 粒度の高いターゲティングが難しいことなどの課題がある. そこで, 筆者らは SNS の一つである Twitter のデータを利用したジオターゲティングに着目した.

Twitter の広告出稿方法にはターゲティング広告の機能が標準で用意されている. 広告出稿者は言語, 性別, 使用プラットフォームや興味関心, 保有するユーザリストなどを指定して広告を出稿できる. この中には地域エリア指定を行う方法も含まれるが, 都道府県レベルの粒度であり, 現実にジオターゲティングを実施するには困難を伴う. そこで, Twitter アカウントの活動履歴から, 滞在するエリアをより細かい粒度で, 高度に予測することができれば, ジオフェンシングを用いるよりも広いカバレッジを確保することができる.

Twitter データを解析することで特定の情報を取り出す方法は盛んに研究されている. Twitter は投稿される SNS の中でも 1 回の投稿が短く, 投稿される回数が多いという特徴を持つ. この特徴を利用し, 実世界で発生したイベントをリアルタイムに検知するソーシャルセンサとして活用する研究がなされている[3, 4]. 同時に, ジオタグ付き Tweet などから位置情報を抽出することで, 特定地域のイベントをリアルタイムに検出することが可能となる.このように, ソーシャルセンサとして利用するため, ユーザの投稿した Tweet からそのアカウントの滞在エリアを推定する手法が検討されている[5, 6, 7]. しかしながら, 粒度に関しては先行研究においても市区町村レベルの試みが限度である.

そこで, 筆者らはより細粒度なTwitterユーザのロケーション分類として, 最大で大字レベルの粒度でのロケーション推定を試みた. 本稿では, 大字レベルでのTwitterユーザロケーション予測の方法を提案する. また, 提案手法に基づいた実装として, 街区名を入力すると当該地点を訪れる可能性の高いユーザのリストを出力するシステムを構築した. 加えて, 構築したシステムを用いたユーザの訪問予測精度について評価実験を実施したので, その結果を報告する.

2 提案手法

本研究ではTwitterユーザの細粒度ロケーション予測，すなわち，あるユーザがある街区を訪れる確率を推定する手法を提案する．本研究で用いるデータの特徴として，ラベル情報が少ないことが挙げられる．これは，ジオタグ付きツイート率の低さに起因する．そこで，提案手法ではラベル情報の拡張を試みる．ここでは，ジオタグ付きツイートを持つユーザの他に，ユーザプロフィールに住所情報を持つユーザを正解ラベル有りのノードとして用いる．次に，ラベル伝搬アルゴリズムの一種であるExpander [8]を用いて住所が既知のユーザを手掛かりに全ユーザの住所を推定する．これに用いるソーシャルグラフの構築手順としては，まず，住所が既知のユーザとフォロー・フォロワー関係にある住所が未知のユーザを隣接ノードとして接続する．次に，住所が未知のユーザ同士をメンション関係で接続する．この状態で，ラベル伝搬アルゴリズムにより正解ラベルを隣接ノードに伝搬させることを繰り返し，住所情報が未知のユーザを含む全ユーザに関する街区を訪れる確率を推定する．

2.1 前提条件

対象となる街区 a は東京都に加えて千葉県，神奈川県，埼玉県，茨城県の隣接4県とし，街区レベル l は都県・市区町村・大字の3段階とした．このとき，レベル l の街区の集合を A_l，全街区の集合 A は $\cup_l A_l$ とする．加えて，街区 a の下位に属するレベルの街区の集合を $A_l(a)$ とする．また，ユーザ u が街区 a を訪れる確率 $p(a|u) \in [0,1]$ のベクトル表記は次式で表される．

$$\mathbf{y}_u = p(a|u)_{a \in A}$$

このとき，住所情報を持つユーザにはあらかじめ，正解確率 $\mathbf{y}_u$ が与えられ，正解ラベルを隣接ノードに伝搬させて，全ユーザの推定確率 $\hat{\mathbf{y}}_u$ を求める．

2.2 ラベル有りノードと正解確率

ラベル有りノードはジオタグ付きTweetと，ユーザプロフィール情報に基づいて設定する．以下に住所情報の付与方法を述べる．

2.2.1 ジオタグ付きTweet

ジオタグ付き日本語 Tweet のユーザごとに投稿地点のクラスタリングを行う．アルゴリズムは DBSCAN を用い，パラメータは ε :1000, minPts:4 とした．また，緯度経度情報からの二地点間の距離計算には hubeny の公式を用いる．得られた結果から，投稿日付の異なり数が 4 以上で投稿日付の数が多い順に 10 クラスタまでを抽出し，クラスタ中心の緯度経度を逆ジオコーディングすることで大字レベルの住所を特定する．

2.2.2 ユーザプロフィール

Twitterユーザの自己紹介/居住地欄に1都4県の街区に該当する文字列を持つユーザを抽出する．街区文字列は都道府県，市区町村，市区に属する大字の名前を対象とする．ただし，ノイズを除去するために秋葉原，六本木，羽田空港といった，曖昧性の高い大字は除外する．抽出したユーザ群に対し，自己紹介/居住地欄の文字列をジオコーディングする．さらに，大字レベルまで特定可能，かつ住所表記に曖昧性がないユーザを抽出する．ジオコーディングにはDAMS[9]を用いた．

2.2.3 正解確率の付与

抽出したそれぞれのラベル有りノードに正解確率を付与する．大字街区 $a \in A_{oaza}$ における正解確率 $p(a|u)$ は，ジオタグ付き Tweet の場合には，訪問日数/tweet 日数，ユーザプロフィールの場合には，経験的に 0.9 を与える．次に，市区町村・都道府県街区 $a \in A_{city} \cup A_{state}$ における正解確率を次式，

$$p(a|u) = \min\left(1.0, \sum_{a' \in A_{oaza}(a)} p(a'|u)\right)$$

によって与える．

2.3 ラベル伝搬

自身を除く全ノードの推定確率 $\hat{\mathbf{Y}}_{-u}$ および正解確率 $\mathbf{Y}$ が所与のとき，ノード損失関数は以下のように定義する．

$$l(\hat{\mathbf{y}}_u; \mathbf{Y}_{-u}, \mathbf{Y}) = \mu_1 \delta_u \left|\left|\hat{\mathbf{y}}_u - \mathbf{y}_u\right|\right|_2^2 +$$

$$\mu_2 \sum_{v \in N(u)} w_v \left|\left|\hat{\mathbf{y}}_u - \hat{\mathbf{y}}_v\right|\right|_2^2 + \mu_3 \left|\left|\hat{\mathbf{y}}_u - \mathbf{c}\right|\right|_2^2$$

ここで，δ_u は u の住所情報の有無(1, 0)，$N(u)$ は u の隣接ノードの集合を意味する．また，$\mathbf{c}$ は一様分布とし，$\mu_i, w_v;\ \sum_v w_v = 1$ はハイパーパラメータである．

$\hat{\mathbf{y}}_u$ の更新式は勾配ゼロの条件より，

$$\hat{\mathbf{y}}_u = (\mu_1 \delta_u \mathbf{y}_u + \mu_2 \sum_{v \in N(u)} w_v \hat{\mathbf{y}}_v + \mu_3 \mathbf{c}) / (\mu_1 + \mu_2 + \mu_3)$$

が得られる．

ここで，$\hat{\mathbf{y}}_u$ に対し，非ゼロ要素の使用メモリを省くため，スパース制約として，更新時に街区レベルごとに上位 N_{K_l} 件の非ゼロ値のみを保存する．また，更新値に街区レベルごとに合計値が 1 となるように正規化を行う．

これらの処理について，各ノードを N_B 件ずつミニバッチ更新し，エポック数 N_I 回繰り返し計算を行う．表1に設定したハイパーパラメータの値を示す．ここで，ウェイト w_v は住所が未知なユーザよりも，既知のユーザからのラベル伝搬を重視するために傾斜をつけている．

Table 1 ラベル伝搬のハイパーパラメータ

Parameter	Value
μ_i	1.0, 0.5, 0.01
N_{K_l}	3, 17, 60
N_B	10^5
N_I	5
w_v	9.0 $if\ \delta_v = 1\ else$ 1.0

Table 4. 東京近郊エリアの予測性能

Address level	@K (K)	Precision @K	Recall @K	F-value @K	Mean AP
都道府県	4,000,000 (1877)	0.731	0.467	0.501	0.761
市区町村	200,000 (94)	0.415	0.129	0.143	0.291
大字	20,000 (10)	0.251	0.056	0.054	0.132

Table 5. 東京都のみの予測性能

Address level	@K (K)	Precision @K	Recall @K	F-value @K	Mean AP
都道府県	4,000,000 (1877)	0.958	0.356	0.519	0.933
市区町村	200,000 (94)	0.500	0.083	0.115	0.348
23 区	200,000 (94)	0.524	0.066	0.103	0.364
大字	20,000 (10)	0.309	0.039	0.044	0.151

Table 6. 大字ごとの予測性能の一例

Area name	Coverage for user	Precision @K	Recall @K	F-value @K	Mean AP
外神田	13,243,238	0.900	0.011	0.021	0.476
丸の内	13,086,760	0.200	0.003	0.005	0.286
永田町	1,080,618	0.100	0.071	0.083	0.076
麹町	1,588,822	0.000	0.000	0.000	0.008

Table 2 データセットおよび登録街区数

Parameter	Value
ラベルありノード	124,454
ラベル無しノード	13,136,541
エッジ数	120 [Mil]
エッジ密度	1.36×10^{-6}
登録街区数	8050
住所情報数	537,883

Table 3 ユーザに対するカバレッジ

Address level	Tokyo	5 states
都道府県	13,256,350	7,956,444
市区町村	3,038,740	772,024
大字	412,225	102,623

3 評価実験

提案手法の有効性を評価するために実験を行った. 提案手法に基づいた実装として, 街区名を入力すると当該地点を訪れる可能性の高いユーザのリストを出力するシステムを構築し, ユーザの予測精度を評価した. 学習用データセットは, 2017~2018 年に東京近郊約 100km 圏内(緯度経度指定)で投稿された Tweet の一部を Gnip Historical Powertrack API を用いて収集した. 表 2 に実際に構築したソーシャルグラフの統計量を示す.

3.1 評価方法

本手法により推定された出現確率 *p(a|u)*は街区 *a* を訪れる確率の高いユーザ *u* を列挙するために使用する. そこで, 定量評価の方法としてランキング学習の評価手法[10]を用いる. 評価指標として, 街区 *a* ごとにユーザ *u* の出現確率 *p(a|u)* の上位 K 件を出力し, その中に正解住所が含まれる割合(精度)を Precision @K, 正解住所のうちにユーザ *u* が含まれる割合(再現率)を Recall @K, およびその調和平均(F 値)を F-value @K として用いる. また, 参考に平均適合率(AP)を示す. 最終的な性能を示す値として, 評価指標を算出した全ての街区における各指標の加重平均を算出する. このとき, 加重平均のウェイトは正解サンプル数n_aに比例させる. @K の水準としては, 全ユーザに外挿した際, ユーザ数が都道府県レベルで 4[Mil], 市区町村レベルで 200[k], 大字レベルで 20[k]となるように調整し, 各街区レベルでそれぞれ, 1877, 94, 10 とした.

3.2 実験結果

収集データのうち, 正解ラベルを持つジオタグ付きユーザの 5%を評価用データとして抽出し, 残りの 95%を学習用データとして評価実験を行った. 表 3 に各街区の訪問確率が非ゼロのユーザ数の平均値を街区レベル別に示す. また, 表 4 に対象範囲を東京近郊とした際の予測性能, 表 5 に東京都に限定した際の予測性能を示す. さらに表 6 に大字ごとの予測性能の一例を示す.

4 考察

実験結果から, 大字＜市区町村＜都道府県といった順で予測精度とカバレッジが高くなっていることが判る. また, 予測精度とカバレッジはどちらも街区ごとにばらつきがあり, 繁華街やターミナル駅といった, 言及されやすい場所を抱える街区の性能が高くなる傾向にあることが判った. これは, エリアマーケティングへの応用という

点で考えると，広告の対象は人の多い地域を対象とする場合が多いため，中吊り広告のように鉄道沿線をターゲットとした広告配信などに応用できる可能性がある．また，都内であれば市区町村レベルで約 300 万のカバレッジと 0.5 程度の予測精度が得られた．東京 23 区の昼間人口は各区において30~80万人程であることから，実店舗を持ち，狭い商圏を対象とする事業者や，折り込み広告，ポスティング・チラシ配布など地域性の高い広告を代替する広告配信に応用できる可能性がある．

次に，提案手法の妥当性について考察する．一般に，コミュニティ検出やノード分類などのネットワーク分析タスクを解決する方法としてネットワークエンベディングにより抽出した特徴量を用いたグラフクラスタリングが主流となっている．しかしながら，文献[13]ではネットワークエンベディングを用いる場合には計算コストとストレージコストの観点から大規模タスクの運用に課題が残ることが指摘されている．また，先行研究[11]によると，フォロー・フォロワー関係による隣接ノードの住所情報が推定に有効であり，加えて，文献[12]では，友人同士（隣接ノード）では住所が類似すると仮定されている．本研究タスクでは低密度なソーシャルグラフを用いて 8000 を超える街区候補を対象とした予測を行う．そのため，提案手法では，隣接ノードの類似性をより直接的に反映できるラベル伝搬法を採用し，さらにメンション関係を用いてエッジ密度を高めたソーシャルグラフを構築している．評価実験の結果として，一定の予測性能を確認できた．その一方で，エッジ情報の拡大・削減・精微化や，ネットワークエンベディングを用いた場合の精度変化の検証は行っていない．この点は今後の課題として確認していきたい．

5 まとめ

本研究では， Twitter 広告を用いたエリアマーケティングを想定し，ラベル伝搬アルゴリズムを応用することで該当エリアに訪れる可能性のある Twitter ユーザを予測する手法を提案した．具体的には，ジオタグ付き Tweet とプロフィール情報から得た住所情報を正解ラベルとして用い，ラベル伝搬アルゴリズムを用いて住所情報が未知のユーザの住所情報を推定し，街区ごとにユーザの訪問確率を予測する．評価実験として，提案手法を約 13 万人の日本語 Twitter ユーザのデータに適用し，1 都 4 県に属する市区町村，および大字を基準とした約 8000 街区を対象にユーザが訪れる可能性の予測を試みた．その結果，東京近郊エリアでの予測精度は，全体を平均して都道府県レベルで 73%，市区町村レベルで42%，大字レベルで25%という結果となった．また，東京都のみでの予測精度は，都道府県レベルで 96%，市区町村レベルで 50%，そのうち 23 区のみだと 52%，大字レベルで 31%という結果となった．

今後の予定としては，提案手法の精度向上を目指し，エッジ情報とラベル情報の拡大・削減・精微化した場合やグラフクラスタリングを用いた場合の推定精度の変化を詳しく検証したい．また，エリアターゲティング機能の実用化に向けて開発を進めていく．実際にサービス化を進めるにあたり，特定の個人を識別することができないよう匿名加工情報として取り扱うことを念頭に，個人情報を加工し，当該個人情報を復元することができないようにした上での利用を徹底していく．

参考文献

[1] Weiqing, W., Hongzhi, Y., Ling, C., Yizhou, S., Shazia, S., Xiaofang, Z.: Geo-SAGE: a geographical sparse additive generative model for spatial item recommendation, ACM KDD, 1255-1264. 2015.

[2] Ankur, G., Sunav, C., Payal, B., Sweta, A., Abhishek, K., Shubham, A.: Smart Geo-fencing with location sensitive product affinity. ACM SIGSPATIAL, 39, 2017.

[3] 榊剛史，松尾豊：ソーシャルセンサとしての Twitter -ソーシャルセンサは物理センサを凌駕するか？-，人工知能学会誌，27(1):67-74. 2012.

[4] Atefeh, F., Khreich, W.: A survey of techniques for event detection in Twitter, Computational Intelligence, 31(1):132-164. 2015.

[5] Cheng, Z., Caverlree, J., and Lee, K.: You are where you tweet: A content-based approach to geo-locating Twitter users, ACM CIKM, 759-768, 2010.

[6] 森國泰平，吉田光男，岡部正幸，梅村恭司：ツイート投稿位置推定のための単語フィルタリング手法，情報処理学会論文誌：TOD，8(4):16-26. 2015

[7] Xin, Z., Jialong, H., Aixin, S.: A Survey of Location Prediction on Twitter, arXiv:1705.03172v2, 24 Feb 2018.

[8] Ravi, S., Diao, Q.: Large Scale Distributed Semi-Supervised Learning Using Streaming approximation, AISTATS, 51:491-499. 2016.

[9] 相良毅，有川正俊，坂内正男：分散位置参照サービス，情報処理学会論文誌，42(12):2928-2940. 2001.

[10] Mcsherry, F., Najork, M.: Computing information retrieval performance measures efficiently in the presence of tied scores. ECIR, 414-421. 2008.

[11] 廣中詩織，吉田光男，岡部正幸，梅村恭司：日本における居住地推定に利用するためのフォロー関係の調査，人工知能学会論文誌， 32(1):1-11. 2017.

[12] Adam, S., Henry, K., Jeffrey. P. B.: Finding your friends and following them to where you are, ACM WSDM, 2012.

[13] Shen, X., Pan, S., Liu, W., Sun, Q. S., Ong, Y. S.: Discrete Network Embedding, IJCAI, 3549-3555. 2018.