4N-03

時間的関係を考慮したラベル伝搬によるツイート発信地推定

上田紗希 * 山口祐人 † 北川博之 ‡ 天笠俊之 ‡

* 筑波大学情報学群情報科学類 † 日本学術振興会特別研究員（PD）
‡ 筑波大学システム情報系

1 序論

本研究では，Twitter に投稿されたツイートの発信地を推定する手法を提案する．ユーザがツイートに付与したジオタグを用いれば，個々のユーザの地域に合わせたニュースや緊急の災害情報などの提供が可能である．しかし，ジオタグは全体のわずか 0.42%のツイートにしか付与されていないため [1]，前述のサービス提供のためには発信地推定が必要であると考えられる．提案手法は，(1) ラベル密度の小さいデータに有効であるラベル伝搬法 [4] を採用し，(2) 同一ユーザが短時間に投稿した複数のツイートは発信地が近いと仮定する．評価実験により，提案手法は既存の推定手法より高い精度で発信地推定が可能であることが示された．

2 関連研究

2.1 位置推定手法

ユーザの居住地を推定する手法は多く提案されている [1][2]．しかし，これらの手法は主にユーザの居住地を推定する手法であり，ジオタグのようなラベル密度の小さいデータに対しては必ずしも有効であるとは言えない．本研究では半教師あり学習の一つであるラベル伝搬法を用いてラベルが少ない問題に対処する．

Ikawa ら [3] は，ユーザの過去のメッセージからツイートの位置情報とキーワードの関連付けを行う手法を提案した．この手法は，教師データのみを用いて発信地推定を行うという点で，本研究とは異なっている．

2.2 ラベル伝搬法

ラベル伝搬法 [4] は半教師あり学習の一つであり，ラベル付きデータ及びラベルなしデータの両方を用いて学習，予測することによりラベル密度が小さい場合にも比較的高い精度を実現する．ラベル伝搬法はまず各データ間の類似度を計算し，その類似度を基に各ノードを各データとする類似度グラフを構築する．そして，構築した類似度グラフを用いてラベルなしノードのラベルを予測する．類似度グラフを図 1 に示す．1，2 はラベルを表し，? はラベルが未知であることを表す．各エッジにはそれぞれ各ノード間の類似度 w_{ij} が対応する．

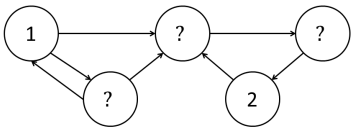

図 1: 類似度グラフ

ノード数を n 個，そのうちラベルが既知のデータが l 個 あるとき，データ間の類似度行列を W とし，その類似度を辺の重みとする重み付きグラフを作成する．n 個 のノードの予測値は，以下の目的関数を最小化する $f = (\mathbf{I} + \lambda\mathbf{L})^{-1}\mathbf{y}$ として求まる．目的関数の 1 項目は予測値と正解値を近づけ，2 項目が隣接したノードの予測値を近づける．

$$J(\mathbf{f}) = \sum_{i=1}^{l} \|y^{(i)} - f^{(i)}\|_2 + \lambda \sum_{i}^{n} \sum_{j}^{n} w_{ij} \|f^{(i)} - f^{(j)}\|_2$$

ただしこのとき，$L = D^{-1/2} W D^{-1/2}$ であり，D は W の各行の和を対角成分に持つ対角行列，$\lambda > 0$ は二つの項のバランスをとるパラメータである．また，y はラベルが既知であるノードのラベルを表す行列である．

3 提案手法

提案手法は，ツイート間の内容の類似度に加えて時間の類似度も考慮する．まず，ツイートから特徴語を抽出し，bag-of-words に従って特徴ベクトルを作成する．特徴ベクトルの各次元の値は tf-idf によって重み付けをする．本研究では内容の類似度 S を計算するにあたり，radial basis function(RBF) カーネルを採用する．また，同一ユーザのツイート間にはそれぞれの投稿時間の近さに応じて時間の類似度 T を定義する．最

Tweet Location Inference via Label Propagation with Temporal Association
Saki UEDA*, Yuto YAMAGUCHI†, Hiroyuki KITAGAWA‡, Toshiyuki AMAGASA‡
*College of Information Science, University of Tsukuba
†JSPS Research Fellow (PD)
‡Faculty of Engineering, Information and Systems, University of Tsukuba

終的なツイート間の類似度 W はパラメータ $\alpha \geq 0$ を用いて $W = S + \alpha T$ と計算する．また，類似度グラフとしては k-Nearest Neighbor(kNN) グラフを採用する．構築した類似度グラフに対してラベル伝搬法を適用し，最も確率の高いラベルをそのツイートの推定ラベルとする．

4 評価実験

4.1 実験設定

提案手法の有効性を検証するため，実際のツイートを用いて発信地推定を行った．ベースラインとしては RBF カーネルを用いた SVM を採用し，提案手法と比較した．さらにパラメータ α の値を変化させ，時間の類似度の効果を検証した．

まず，Twitter Streaming API を用いて 2014 年 9 月 2 日から 2014 年 11 月 20 日の間に日本のジオタグ付きツイート 3,526,922 件を収集した．次に時間の類似度の効果を確かめるため，ジオタグ付きツイートを複数投稿しているユーザの時間的に近いツイートを収集した．取得したツイート中に 2 回以上出現するユーザからランダムに 1 万人を選出し，さらにそのユーザらの直近 200 ツイートからジオタグ付きツイートを 170 件以上含んでいるユーザのジオタグ付きツイートを抽出した．最終的にユーザは 707 人，ツイート総数は 130,705 件であった．これらをトレーニングデータ，バリデーションデータ，テストデータに分割し，トレーニングデータとバリデーションデータで適切なパラメータを決定した．

また，調整したパラメータは提案手法の λ，RBF カーネルの γ，kNN グラフの k，SVM のパラメータ γ と C である．その後，トレーニングデータとバリデーションデータを合わせてラベル付きデータとし，テストデータのジオタグを隠して発信地推定を行い正解率を計算した．今回は特徴語として地名を用い，形態素解析には MeCab[1] を使用した．推定される発信地のラベルは経度 122 度から 145 度，緯度 23 度から 45 度の地図を 1 度ずつのグリッドに区切った全 552 ラベルとした．

4.2 実験結果と考察

提案手法についてパラメータ α の値を変化させた場合の正解率と SVM との比較を表 1 に示す．ラベル付きデータの割合は 0.5%，時間の類似度 T は，同一ユーザの 60 分以内のツイート間については 1，それ以外は 0 とした．また，$\alpha = 2147483647$ は 32bit 整数の最大値である．

時間の類似度を導入することで，提案手法の正解率は向上していることが分かる．これは，同一ユーザの時間的に近いツイートは地理的に近い地点から投稿されていたため，内容の類似度のみでは正しく推定されなかったツイートを正しく推定することができたからだと考えられる．さらにパラメータ α の値を調整することで，同じラベル付きデータの割合の SVM よりも高い精度で推定を行うことができた．

表 1: 正解率

S, T 利用	α の値			
	0	1.0	10	2147483647
	26.27%	27.12%	**28.38%**	27.99%
T のみ	27.21%			
SVM	26.87%			

5 結論

本研究では，ラベル伝搬法を用いてツイート内容と投稿時間からそのツイートの発信地を推定する手法を提案した．これによって，発信地推定におけるラベル密度が小さい問題に対処する．今後の課題として，さらに有効な特徴語の抽出方法や時間の類似度の計算方法の改善の検討が挙げられる．

謝辞

本研究の一部は，文部科学省「実社会ビック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．

参考文献

[1] Z.Cheng, J.Caverlee, and K.Lee. ”You are where you tweet: a content-based approach to geo-locating twitter users.” In Proc. CIKM 2010, pp.759-768, 2010.

[2] Backstrom, L., Sun, E. and Marlow, C. ”Find Me If You Can: Improving Geographical Prediction with Social and Spatial Proximity.” WWW, pp.61-70, 2010.

[3] Y. Ikawa., M. Enoki., M. Tatsubori. ”Location Inference using Microblog Messages.” WWW, pp.687-690, 2012.

[4] Zhou, D., Bousquet, O., Lal, T. N., Weston J., and Sch ¨ olkopf B. ”Learning with Local and Global Consistency.” NIPS 16, pp.321-328, 2004.

[1] http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html