M-030

位置情報付き投稿におけるテキスト解析を用いたラベル付け手法の検討
A Labeling Method using Text Analysis on Geotagged of Microblog

酒巻　智宏 [1]　　岩井　将行 [2]　　瀬崎　薫 [3]
Tomohiro SAKAMAKI[1]　　Masayuki IWAI[2]　　Kaoru SEZAKI[3]
東京大学大学院新領域創成科学研究科 [1]　　東京大学大学院生産技術研究所 [2]
東京大学空間情報科学研究センター [3]

1　背景と目的

住宅から職場へ人々がどのように移動しているか，といった人の行動を知ることは，マーケティングや交通などの需要予測を行う際に重要な情報である．現在，人の行動を把握するために，アンケート調査 [1]，GPS を用いた手法 [2] ，携帯電話基地局の通話履歴情報を用いた手法 [3] など，様々なアプローチが取られている．その中で，本研究ではマイクロブログ，特に最も利用者の多い Twitter に注目する．Twitter は現在の自身の状況を短いメッセージ (Tweet) で表現することでコミュニケーションを行う SNS サービスである．Twitter の機能として，投稿にユーザの現在位置情報を付加することができる“ ジオタグ ”というサービスがある．

位置情報とユーザの現在の状況を含むジオタグ付き Tweet は，ユーザの現実世界での行動履歴の性質を持っている．そこで，ジオタグ付き Tweet からその位置がそのユーザにとってどのような意味を持つか推定できると考えられる．例えば,「会議始まった」というジオタグ付き Tweet があれば，その地点はそのユーザにとって仕事に関連する場所であると推定できる．

我々の研究の目標は，ジオタグ付き Tweet を用いて，人の行動調査を行うことである．これまで我々は，クラスタリングを用いて人が日常的に滞在している場所を発見する研究を行った [4]．本稿ではさらに，個々のユーザが日常的に滞在している場所に対して家や仕事場のように意味付けを行うことを目標とする．

2　関連研究

1 章でも述べたように，近年様々な手法を用いて人の行動の調査が行われている．GPS 付き携帯電話端末を用いて人の行動調査を行う研究 [2] では，GPS 情報を用いることにより，リアルタイムに，かつ従来型のアンケート調査による行動調査と比較して作業負担の軽減を可能にしている．

また，携帯電話基地局の通信履歴を用いる研究 [3] では，ユーザが頻繁に通信する携帯電話の基地局をクラスタリングすることで人の行動の特徴点の発見を行い，加えてロジスティック回帰を用いて家や仕事場というラベル付けを行っている．

位置に対してユーザがどのようなラベル付けを行うかを分析した研究 [5] では，被験者に GPS を持たせ行動をトレースし，自分がよく活動する場所に対して任意にラベルを付けてもらい，その傾向の分析とラベル付けのパターンについての適切な分類を行った．この研究では，最終的に，自動で位置に対してラベリングを行うことを目標としている．しかし，この目標を達成するためには，位置情報だけでは不十分であると指摘している．

3　提案手法

3.1　クラスタリングによる特徴点の発見

まず，前回の研究と同様，ジオタグ付き Tweet T を収集し，T の lat, lon 情報によりクラスタリングを行う．群平均法をベースとして以下の手順でクラスタ C を抽出する．

1. あるユーザがこれまでに投稿したジオタグ付き Tweet T を収集する
2. 各 T の lat, lon をもとに，群平均法を用いてクラスタリングを行う
3. もしあるクラスタ C_i の作る円の半径 r が 1km を超えたら 4 へ
4. C_i に含まれる Tweet 群 $T_C = \{T_{C1}, \cdots, T_{Ck}\}$ が以下の 2 つの条件を満たす場合は，そのクラスタを採用し，2 へ戻る．そうでなければそのクラスタを除外し，2 へ戻る
   - そのユーザの全 Tweet 数の 1%以上
   - $T_{C1}, \cdots, T_{Ck}$ の取得日 dt が 5 日以上分散されている

図 1 は，上記の手法を用いて実際にクラスタリングを行った例である．図の小さいアイコンが各ジオタグ付き Tweet を，大きいアイコンがクラスタの中心点を表している．

3.2　テキスト解析によるクラスタへの意味付け

次に，クラスタリングにより取得された各クラスタに対して，家や仕事場といったラベル付けを行い，その場所がそのユーザにとってどのような意味を持つ場所かを推定する．ユーザは，場所に応じて様々な内容の Tweet を投稿していると考えられる．例えば，食事を行う場所であれば，食事の感想を投稿し，職場では仕事に関する投稿をする可能性が高い．そこで，クラスタ内の Tweet T に含まれる単語を用いることでクラスタに対してラベル付けを行う．

図 1: クラスタリング結果の例

本研究では，クラスタに含まれる Tweet 群をひとつのドキュメントと仮定し，ドキュメント分類によく利用される Naive Bayes を適用し，クラスタを以下の 6 つに分類する（表 1）.

表 1: ラベル一覧

	ラベル一覧
家	ユーザの自宅
仕事場	ユーザの職場
学校	ユーザの通う学校
娯楽	買い物や食事など，余暇を過ごす場所
移動	鉄道や飛行場など交通機関に関する場所
その他	その他分類外の場所

単語の抽出は，以下の手順で行う.

まず，各クラスタ内の投稿内容 t から，RT や@付き投稿など Twitter 特有の表現記法を消去した後，形態素解析を行い，名詞・形容詞・動詞・副詞に該当する単語を抽出する．なお，形態素解析には MeCab * を用い，単語辞書には IPA 辞書 † を用いる.

4 評価実験と考察

4.1 評価実験

提案手法の有効性を検証するために，日常的にジオタグを利用している 15 人の Twitter ユーザに被験者として協力していただき，ジオタグ付き Tweet を用いて評価実験を行った.

まず，被験者のジオタグ付き Tweet を元に，3.1 節で示したクラスタリングによりクラスタを取得した．次に，取得したクラスタに対して，手動で表 1 の 6 種類のラベル付けを行った.

その後，3.2 節で示した Naive Bayes を用いた手法により，各クラスタに含まれる Tweet 内容を解析し 6 つのラベル分類を行った．評価は 15 人のサンプルデータのうち 1 人をテストデータとして取り出し，残りを学習データとする leave-one-out cross-validation により行った．15 人全員について手動で付けたラベルと分類器によるラベルを比較し，平均の適合率を算出した.

4.2 実験結果

表 2: cross-validation の結果

前被験者から獲得したクラスタの総数	150
適合率	0.613

実験結果は表 2 で示すとおりである．15 人の被験者から獲得した全クラスタ数は 150 であった．これらのクラスタに手動でラベル付けを行い，cross-validation を行った結果，適合率は 0.61333 となった.

5 考察と今後の課題

今回の評価実験では適合率が 0.613 と良い精度が出なかったが，理由として，ユーザの Tweet は自身の行動以外にも思考やニュースの引用なども含まれることが主な原因として考えられる．より精度を上げるためには Tweet のうち，ユーザの行動のみを分類して利用する方法が考えられる.

また，クラスタリングで誤った位置にクラスタが生成されてしまう例が見られたが，実験データを見ると，被験者が鉄道などで移動中に投稿した Tweet からクラスタが生成されるためであることがわかった．今後の課題として，頻繁に利用する交通機関の情報を抽出する場合には，円状ではなく線上に分布した Tweet 群を抽出する手法が必要になるとわかった.

6 まとめ

本研究では，ジオタグ付き Tweet を用いた人の行動調査について，ユーザが頻繁に行動する位置に対しての意味付けを，Naive Bayes を用いたテキスト解析により行った.

参考文献

[1] 森尾淳, 中野敦. パーソントリップ調査の実態調査上の問題点と改善手法. *IBS Annual Report*, pp. 86–88, 2006.

[2] 松本修一. Gps 携帯を活用した行動調査に関する基礎的研究. *KEIO SFC JOURNAL*, Vol. 9, No. 1, 2009.

[3] Sibren Isaacman, Richard Becker, Stephen Kobourov3, Margaret Martonosi, James Rowland, and Alexander Varshavsky. Identifying important places in people's lives from cellular network data. *Pervasive Computing*, Vol. 6696, pp. 133–151, 2011.

[4] 酒巻智宏, 岩井将行, 瀬崎薫. マイクロブログのジオタグを用いたユーザの行動パターンの推定に関する研究. 第 2 回集合知シンポジウム, 2010.

[5] Jialiu Lin, Guang Xiang, Jason I. Hong, and Norman Sadeh. Modeling people's place naming preferences in location sharing. *Ubicomp '10*, pp. 75–84, 2010.

*http://mecab.sourceforge.net/
†http://sourceforge.jp/projects/ipadic/