マイクロブログのジオタグを用いたユーザの行動分析

Analysis of Human Behavior of a User using Geotag of Microblog

学籍番号　47-106756
氏　名　酒巻 智宏（Tomohiro, Sakamaki）
指導教員　瀬崎 薫 教授

1.背景と目的

Web は元来，現実世界とは無関係のバーチャルワールドであると考えられてきた．しかし，SNS やブログ，更には自身の状況をリアルタイムに投稿するマイクロブログの登場により，両者の距離は急速に縮まった．その中で，マイクロブログから現実世界の情報を抽出する手法に注目が集まっている. 本論文では, このような背景の中で, 全世界で最もユーザ数の多いマイクロブログである Twitter に注目し, Twitter からユーザの現実世界での行動情報を抽出する手法について研究を行う．

本研究では，Twitter のジオタグという位置情報を付加した投稿(Tweet)に注目し，ジオタグ付き Tweet の属性を用いたクラスタリングとラベリングによってユーザの行動を把握することを目的とする．具体的に，ユーザが「定期的にどこで活動しているか」「その場所でどのような活動をしているか」の２点について解析を行う．

ただし，本研究の仮定として Tweet と現実世界の行動が一致していることを前提としている．また，行動分析の対象は Twitter 利用者のみに限定される．

2.関連研究

2.1.ジオタグを用いたデータマイニング

ジオタグを用いた Tweet 解析は，Twitter 全体から情報を取得するものが多い．例えば，イベント検知を行う研究[1]があげられる．これに対し，本研究では個人をターゲットとしてデータマイニングを行う．

2.2.行動調査に関する研究

人の行動について把握するために，一般的には国勢調査が行われている．近年，国勢調査に変わり，携帯電話基地局情報を用いてユーザの行動調査を行う試みがなされている[2]．携帯電話基地局を用いる研究は多数あるが，本研究のように行動の意図まで考慮している研究は少ない．

2.3.位置に対するラベリング関する研究

位置に対するラベリングに関する研究[3]では，被験者が頻繁に訪れる場所に対して自動で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．そこで本研究では，その位置で投稿されている Tweet 内容に注目することで，位置に対して意味付けを行う．

3.提案手法

まず，ユーザのジオタグ付き Tweet の位置情報，時刻情報，投稿内容を用いクラスタリングを行い，ユーザがよく活動していると思われる地点のクラスタを抽出する．

次に，各クラスタ内の投稿内容を単語ベクトル化し，ナイーブベイズによって家や

職場などに分類し，ユーザにとってその場所がどのような場所であるかを推測する．

3.1.クラスタリングによる特徴点の発見

本研究では，クラスタリング手法として Newman 法を用いる．Newman 法は凝集法の一種であり，密に結合したノードのまとまりを発見するのに適した手法である．ジオタグ付き Tweet はユーザの活動する場所に密に分布することが多く，Newman 法が適していると考えられる．

Newman 法はノード間に重みを付けることができる．そこで，ジオタグ付き Tweet をひとつのノードとし，ノード間の距離，時間差，投稿内容の類似度によりノード間の重みを算出し，クラスタリングを行う．

まず緯度経度から得られる距離を重み付けに利用する．本研究では，ノード n(i), n(j) に対して以下のような式で重み付けを行う．

$$w_l(i,j)=\begin{cases}1-\dfrac{|n_l(i)-n_l(j)|}{L} & (|n_l(i)-n_l(j)|\le L)\\ 0 & (|n_l(i)-n_l(j)|> L)\end{cases} \quad (1)$$

距離が遠いノード同士は無関係であると考えられるため, L = 1000(m) 以上のノードについては重み付けを 0 とする．

また，ユーザは，ひとつの場所には同じような時間帯に滞在することが多い．時間の要素を追加することでより実際の活動に即したクラスタの抽出を行う．Tweet の投稿時刻を元に以下の式でノード同士の重み付けを行う．

$$w_t(i,j)=\begin{cases}1-\dfrac{|n_t(i)-n_t(j)|}{T} & (|n_t(i)-n_t(j)|\le T)\\ 0 & (|n_t(i)-n_t(j)|> T)\end{cases} \quad (2)$$

ただし，時間差を算出する際は日付情報を用いず単純に時刻を比較して計算を行う．例えば，2011/08/31 12:00:00 と 2011/08/29 11:00:00 の差は 60 分となる．本研究では，T = 180(分)とし，それ以上離れた 2 点については重みを 0 とする．

さらに，投稿内容の類似度もクラスタリングの重み付けとして用いる．本研究では，投稿内容に含まれる単語を単語ベクトル化し，コサイン距離を取ることでノード同士の重み付けを行う．ここで，N(i), N(j) はノード n(i), n(j) の単語ベクトルである．

$$w_c(i,j)=\frac{N(i)N(j)}{\sqrt{|N(i)||N(j)|}} \quad (3)$$

これらをまとめると，ノード i, j 間の重み付けの式は以下で表すことができる．

$$w(i,j)=w_l+\alpha w_t+\beta w_c \quad (4)$$

これら α, β の値については，評価実験を行うことで適切な値について考察する．

3.2.ラベリングによる意味付け

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，食事の感想を投稿する可能性が高い．そこで，クラスタ内の Tweet 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．

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 Naïve Bayes, Complement Naïve Bayes を適用し，クラスタを以下の 6 つに分類する（表 1）．単語の抽出は，まず各クラスタ内の投稿内容から RT や@付き投稿など Twitter 特有の表現記

法を消去した後，名詞・形容詞・動詞・副詞に該当する単語を抽出することで行う．

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	余暇を過ごす場所
移動	交通に関する場所
その他	その他分類外

4.評価実験

本研究の手法について評価実験を行った．実験は，日常的に Twitter とジオタグを利用している 25 人の被験者に対して行った．事前に被験者が日常的に訪問している場所について，被験者に対してアンケートを行い，自宅や勤務先など，被験者が訪問している場所について任意の数だけ回答を得た．

4.1. クラスタリングに関する評価実験

3 章で示したクラスタリング手法を用いてユーザの行動について分析した．クラスタリングは，式 4 の各変数の値を変えながら様々なパターンで行った．クラスタリングには被験者の最新 1000 件のジオタグ付き Tweet を利用した．取得されたクラスタの中心点とアンケートによる位置との物理的な距離の総和 D を比較することで，適切なクラスタリングが行えているかを評価した．取得されたクラスタは，クラスタにふくまれるジオタグ付き Tweet の数で並び替え，上位からアンケートにより取得された地点と同数だけ利用する．評価はすべてのユーザの平均値により行う．

Pattern A：まず，式 4 において，$\alpha = 0, \beta = 0$ という条件でクラスタリングを行った．つまり，緯度経度距離のみを用いる条件である．この条件では，平均誤差距離 D は 1893.5m という結果になった．

Pattern B：次に，$\alpha = x, \beta = 0$ という条件でクラスタリングを行った．つまり，時空間情報を用いる条件である．x を変化させて計算を繰り返し，結果を図 1 にまとめた．

この場合，$\alpha = 0.01$ の時，D が 1736.9m と最も距離の誤差が短くなった．

Pattern C：次に，$\alpha = 0, \beta = x$ という条件でクラスタリングを行った．つまり，緯度経度距離とテキスト情報を用いる条件である．x を変化させて計算を繰り返し，結果を図 2 にまとめた．この場合，$\beta = 0.1$ の時 D が 1748.9m と最も距離の誤差が短くなった．

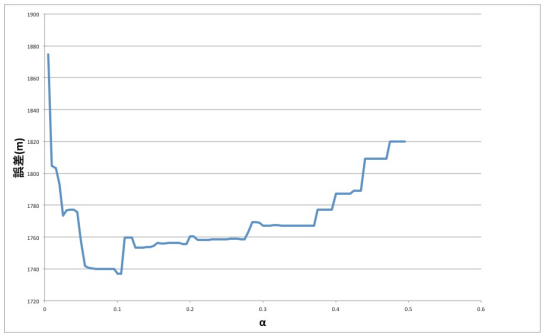

図 1: Pattern B の結果

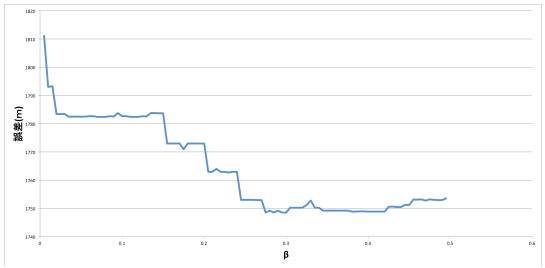

図 2: Pattern C の結果

Pattern D：最後に，すべての要素を含めた条件として，$\alpha = x, \beta = y$ という条件でクラスタリングを行った．ここでは，x をそれぞれ 0.1, 0.2, 0.3 とし，y は 0 から 0.5 までそれぞれ 0.01 刻みで変化させ，計算を行った．この条件では，$\alpha = 0.1, \beta = 0.3$ の時，D は 1665.7m という結果になった．

実験の結果を表 2 にまとめる．実験では，Pattern D が最もごさが少ないという結果になった．

表 2: 実験結果

	D(meter)
Pattern A	1893.5
Pattern B(最良値)	1736.9
Pattern C(最良値)	1748.9
Pattern D	1665.7

4.2. ラベリングに関する評価実験

上記の実験で取得されたクラスタに対して，3.3 章で示したラベリング手法を適用した．まず，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，提案手法により各クラスタに含まれる Tweet を解析しラベル分類を行った．評価は 25 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. 適合率は Naïve Bayes を用いた場合 0.643，Complement Naïve Bayes を用いた場合 0.696 となった．

5.考察

今回の評価実験で得られた結果について考察する．クラスタリングについて，最も精度が良かったのは位置，時間，テキストすべてを利用した Pattern D であったが，位置＋時間，位置＋テキストを用いた Pattern B, Pattern C もほとんど同程度の誤差でクラスタを抽出できていた．このようなデータセットをクラスタリングする際には，位置のみでなく時間や投稿内容を利用することが有効であると言える．

ラベル付けに関して，評価実験では適合率が 0.613 と良い精度が出なかった．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．精度を上げるためには，Tweet のうちユーザの行動に関するものを分類して利用する方法が考えられる．

6.まとめと今後の課題

本研究では，ジオタグ付き Tweet を用いてユーザの行動情報について分析を行った．ジオタグ付き Tweet の属性を利用したクラスタリングついて考察を行い，位置情報だけでなく，時間やテキストを利用することで，より精度の良いクラスタリングができることを確認した．

さらに，クラスタにラベル付けを行い，その場所でのユーザの活動を把握する方法を提案した．評価実験ではあまり良い精度で分類できなかったが, ユーザの Twitter 利用パターンを分析することで，より精度を上げることができると考えられる．

今後の課題としては，クラスタリング・ラベリングの精度向上と本研究で抽出した行動情報を利用するアプリケーションの開発を行いたい．

文　献

[1] R. Lee et al. Measuring geographical regularities of crowd behaviors for twitter based geo-social event detection. LBSN 2010.

[2] S. Isaacman et al. Identifying important places in people's lives from cellular network data. Pervasive Computing 2011.

[3] J. Lin et al. Modeling people's place naming preferences in location sharing. Ubicomp 2010.

発　表　歴

情報処理学会 第 73 回全国大会 （2011.01）
第 2 回集合知シンポジウム （2011.03)
FIT2011 (2011.09)
ICHPSS 2011 (2011.12)