1H3-4in

位置情報付きツイートから抽出した交通路の評価

Evaluation of Traffic Route Extraction from Geotagged Tweets

谷 直樹 *1 Naoki TANI　風間 一洋 *1 Kazuhiro KAZAMA　榊 剛史 *2 Takeshi SAKAKI　吉田 光男 *3 Mitsuo YOSHIDA

*1和歌山大学 Wakayama University　*2株式会社 ホットリンク Hotto Link Inc.　*3豊橋技術科学大学 Toyohashi University of Technology

In order to discover a variety of routes that are dynamically generated under certain circumstances, such as "Toreta Road Map" which shows safety car routes estimated from probe car data after catastrophic earthquakes, we propose a method to estimate traffic routes, which a certain type of users passed through, from geotagged tweets. We applied the method to traffic route extraction of public transportation on which many passengers are carried routinely. Furthermore, we evaluate recall calculated by root mean squared errors between our estimated route and actual routes that are derived from railway data of the Ministry of Land, Infrastructure, Transport and Tourism in Japan.

1. はじめに

スマートフォンの普及に伴い，身の回りで起きた出来事を簡単に発信できる Twitter が注目されている．特に位置情報が付加されたジオタグ付きツイートは，ユーザの発言場所や移動経路がわかるだけでなく，発言内容と組み合わせて分析することで，例えば地震の震源地や台風の移動経路などの現実世界の状況を把握するためのソーシャルセンサーとして活用できる [1].

我々は，ジオタグ付きツイート中で，ある同一条件を満たすユーザが共通で利用した移動経路を推定する手法を提案し，実際に日常的に多くの利用者を運ぶ公共交通機関の交通路の抽出に適用した [2]．このような公共交通機関の経路データは他の手段でも入手できるが，東日本大震災時の「通れたマップ」のような震災直後に車が通行できた経路や，直後の停電や安全確保のための交通機関の運行停止による帰宅難民の行動の把握，または花見や紅葉の時期の名所の把握など，ある状況下で動的に生成される様々な経路の発見に適用できると考えられる．

ただし，ユーザの散発的なツイートでは経路を忠実に再現するには不十分であり，ユーザの位置しか取得できないために交通機関の経路だけを切り出せないことから，既存の経路抽出法をそのまま適用できない．そこで，投稿中又は前後に高速な交通手段を利用したと思われるツイートを抽出し，対象区域を細分化した各矩形領域内で近接している二つのツイートを Hough 変換することで交通路の断片と思われる近似直線を求め，それらをグループ化することで多くのユーザが利用した公共交通機関の交通路を抽出する．

さらに，国土交通省が提供している国土数値情報鉄道時系列データに含まれる実路線の地点リストと，提案手法で求めた近似直線の間の距離と角度の平方根平均二乗誤差を求めて，閾値を超えた場合に正確に抽出できたと考えて，JR 東日本の山手線全線に対して再現率を求める．また，一部の再現性の悪い場所については，Google Maps 上の近似直線と地点リストの可視化結果に基づいて原因を考察し，本手法の性能改善方法について考察する．

連絡先: 谷 直樹 (s161032@sys.wakayama-u.ac.jp)
和歌山大学システム工学部
〒 640–8510 和歌山県和歌山市栄谷 930

2. ジオタグ付きツイートからの交通路の抽出

2.1 ジオタグ付きツイートを用いる場合の問題点

例えば，東日本大震災直後に車が安全に通行できた経路を示したサービスである「通れたマップ」は，Hada らの GPS と通信機能を搭載したプローブカーを用いて収集した情報を分析する研究 [3] に基づいている．しかし，本稿のようにジオタグ付きツイートから交通路を抽出する場合には，これらの既存研究にない次の問題点が存在する．

1. 位置取得タイミングの制御の問題．位置の取得タイミングはユーザのツイート行動に依存する．通常は取得間隔が長い上に，タイミングも不定なので，個々のユーザのジオタグ付きツイートだけから交通路を再現できない．
2. 移動手段の位置取得の問題．交通機関ではなくユーザの位置しか取得できないので，常に交通路上にあるとは限らない．
3. ユーザの移動手段利用判定の問題．連続する二つのツイート間の移動速度から移動の有無は推定できても，ツイートした瞬間に移動していたかどうかは推定できない．

そこで，以下の手順で構成される交通路抽出法を用いる．

2.2 ボットアカウントの除外

Twitter ボットは自動的にツイートするプログラムであり，設定した文章を自動でツイートするボット生成サービスも多く利用されている．ジオタグを付加してツイートするボットも存在するために，人間のつぶやきだけを取り出すために，分析前にボットアカウントを除去する．

一般的に，利用クライアント名（source 値），ユーザ名（screen_name 値），プロフィール情報（description 値）でボットアカウントであることを明示することが多いので，これらに「BOT」,「Bot」,「bot」などの単語が含まれている場合は処理対象から除外する．

2.3 移動ツイートの抽出

ジオタグ付きツイートからユーザの位置は取得できても，その時に自動車，バス，電車，新幹線などの交通機関に乗車していたかどうかは判別できない．代わりに，ユーザの連続する二つのツイートの投稿位置と時間から求めた平均移動速度が閾値 T_v 以上の場合に，二つのツイートの間に上記の交通機関で

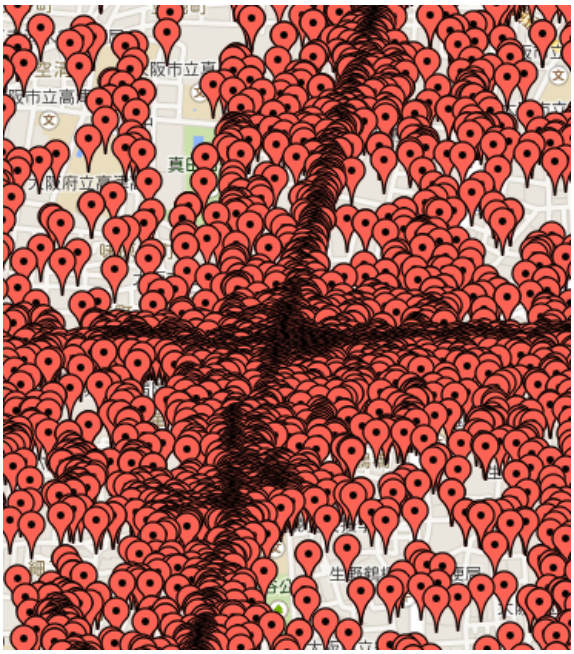

図 1: 鶴橋駅周辺の移動ツイートの可視化例

移動したとみなす．なお，ここで対として扱うのは，前後に交通機関で移動したツイートを抽出するためだけであり，その後の処理では再び分解して独立したツイートとして扱う．本稿では，この処理で得られたツイートを**移動ツイート**と呼ぶ．

ただし，移動ツイート集合がそのまま交通路を表すわけではない．例えば，近鉄大阪線と JR 大阪環状線が交差する鶴橋駅周辺の移動ツイートを Google Maps 上に可視化した結果を，図 1 に示す．黒いエッジ部を持つマーカの密集で黒い部分が生じるが，確かに近鉄大阪線と JR 大阪環状線に対応する縦と横の黒い軌跡が観測でき，中央の交差地点が鶴橋駅である．しかし，近鉄大阪線の左部分では移動ツイート量の不足により黒い部分が観測されず，逆に図の中央下部の路線と関係のない部分に黒い部分が観測される．さらに，主要な繁華街・オフィス街や複数路線の乗り換え駅では，目視による判別が不可能なほど大量の移動ツイートが広範囲に渡って密集する．

2.4 交通路の近似直線の抽出

次に，対象地理空間を複数の矩形領域に分割し，各矩形領域内の移動ツイート群の特に密度が高い連続部分を **Hough 変換** [4] を用いて，交通路を部分的に近似する直線を抽出する．

Hough 変換は，画像から得られた多くのエッジを，原点から垂直に引いた直線の距離と角度の空間（Hough 空間）上に写像し，パラメータ頻度が高い箇所を再び元の空間上に逆写像することで，エッジ群を通る直線を抽出する手法である．原点からの距離を ρ，角度を θ とすると，以下の式 1 が成り立つ．

$$\rho = x\cos\theta + y\sin\theta \tag{1}$$

ただし，移動ツイートでは，すでに述べたように前後に移動したかどうかを推定できるだけで，必ずしも交通路上にあるとは限らない．例えば，自宅でツイートしてから電車で移動し，待ち合わせをしていたレストランで再びツイートした場合には，どちらのツイートの位置も交通路とは離れた場所となる．このような場合は移動ツイート対を繋ぐエッジは交通経路とはかけ離れた位置・角度を持つために，Hough 変換の元データとしてそのまま使うことはできない．

そこで，時系列的な連続性は無視し，距離的に近接する二つの移動ツイートを繋いでエッジとすることで得られるエッジ集合を入力データとする．この二つの移動ツイートは，必ずしも同一ユーザではなく，エッジも交通路上に乗っているとは限らないために，通常の Hough 変換と違って入力に大量のノイズが混在しているが，パラメータ頻度が高い箇所だけを逆変換することで，ノイズが除去する．

各矩形領域内の近似直線を抽出するアルゴリズムは以下の通りである．

1. 閾値 T_d 以下の距離の二つのツイートを通過するエッジ群を求める．
2. エッジ群を Hough 変換し，Hough 空間上の距離・角度に対して分割された領域ごとのエッジ頻度を集計する．
3. Hough 空間上のエッジ数が最大となる領域の距離，角度の平均値を求めて，地理空間上に直線として逆写像する．
4. 矩形領域内のユーザ数と集中度が，閾値 T_u と T_c を下回る場合は，ノイズとみなして処理対象から除外する．

なお，矩形領域内のユーザ数がユーザ数が少ない領域は特定ユーザの自宅や職場である可能性が高いので，閾値 T_u を下回る場合には処理対象から除外する．

さらに矩形領域 (i,j) 内の分布が特定部分に偏っているかどうかを表す指標である集中度 $c_{i,j}$ を，移動ツイート数 $t_{i,j}$ とエッジ数 $e_{i,j}$ を用いて，以下のように定義する．

$$c_{i,j} = \frac{e_{i,j}}{t_{i,j}} \tag{2}$$

集中度は移動ツイートが密集しているほど大きくなり，均一に分散している場合は小さくなる性質を持つことから，閾値 T_c を超える場合のみを対象とする．

2.5 同一経路と推定される近似直線のグループ化

上記の処理で独立した短い直線群が抽出されるが，電車の路線などの経路は連続した長い直線群であることから，さらに同一経路上にあると思われる近似直線をグループ化する．

矩形領域 (x,y) の近似直線 l と，同一経路上にあると推測される近似直線 l' を発見する概念図を図 2 に示す．ここで，l と l' の中線を結ぶ線分を引き，l と水平軸のなす角度を α，l' の角度を β，中線の角度を γ とする．これらの角度が，電車や道路などの曲率を考慮した範囲差であれば，近似直線 l と l' は同一経路集合に属するとする．

ただし，GPS による位置計測や通信に支障があるなどの理由で位置が取得できない又は不正確だったり，乗客数が少ないなどの理由で充分なデータが得られない場所もあると考えられる．そこで，グループ化時にある程度のギャップは許容するために，矩形領域 (x,y) の近似直線 l の相手の近似直線 l' の探索範囲を $(x-2,y-2),(x+2,y-2),(x-2,y+2),(x+2,y+2)$ の 24 個の矩形領域とする．これにより，例えば隣接 8 矩形領域に近似直線が見つからない場合でも，その先の 16 矩形領域に近似直線が存在すればグループ化できる．

具体的なアルゴリズムを以下に示す．

1. 指定された範囲から，未処理の近似直線 l' を一つ取り出す．未処理の近似直線がない場合は終了する．
2. $diff(\alpha,\beta) \leq T_{\theta_1}$ を満たさなければ，(1) に戻る．
3. $diff(\alpha,\gamma) \leq T_{\theta_2} \wedge diff(\beta,\gamma) \leq T_{\theta_2}$ を満たさなければ，(1) に戻る．
4. 直線 l' を同一経路グループに追加して，(1) に戻る．

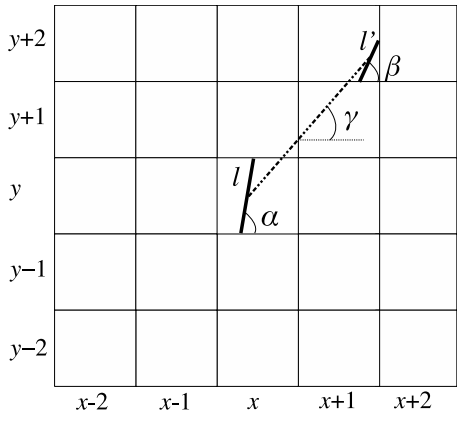

図 2: 近似直線同一経路集合を構築する例

なお，l の角度を α，l' の角度を β，l と l' の中線の角度を γ，$diff(\theta, \theta')$ は，角度 θ，θ' の角度差 $\theta''(0 \leq \theta'' < 360)$ を求める関数，T_{θ_1}，T_{θ_2} は角度の閾値とする．抽出された近似直線にさらに同じアルゴリズムを提供することで，同一経路と思われる近似直線グループが抽出できる．

なお，近似直線グループサイズが閾値 T_g より小さい場合には，明確な交通経路ではないノイズと考えて除外する．

3. 評価

3.1 ツイートデータセット

Twitter Streaming API[*1] を用いて，ジオタグが付いたツイートだけを収集し，JSON 形式で保存したツイートデータセットを作成した．この中から，2013 年 11 月 1 日〜4 月 31 日の 6ヶ月分を抽出して，評価に使用した．データに含まれるツイート数は 67,619,243，ユーザ数は 1,130,328 である．

3.2 パラメータ

関東の JR 山手線周辺区域に対して交通路抽出を行った．領域の面積は，JR 山手線周辺区域が 200km^2 であった．

移動ツイート対の抽出における速度の閾値 T_v は，想定した移動手段で最低速度と考えられる各駅停車の速度である 18km/h とした．実際には各駅停車の移動速度は 24km/h の区間が多かったが，18km/h 区間も対象にできるように，低い値を閾値とした．この結果，JR 山手線周辺区域から 438,175，JR 大阪環状線周辺区域から 156,450 の移動ツイートが抽出された．

交通路の近似直線の抽出に関しては，対象領域は 250m の矩形領域に分割し，Hough 空間は，距離の軸では 100m，角度の軸では 10 度ごとに分割した．

さらに，ツイート間距離の閾値 T_d を 100m，ユーザ数の閾値を T_u を 3，集中度の閾値 T_c を 0.7 とした．

同一経路と推定される近似直線のグループ化においては，角度差の閾値 T_{θ_1} と T_{θ_2} を共に 30 度とした．さらに，同一経路近似直線グループのサイズの閾値 T_g を 3 とした．

3.3 評価用路線データ

抽出結果を評価する際に正解とする路線データとして，国土交通省が全国総合開発計画，国土利用計画，国土形成計画などの国土計画の策定や実施の支援のために作成した国土数値情報の中から，鉄道時系列データ [*2] を用いた．鉄道時系列データは，XML ベースのマークアップ言語である GML を用いた地理情報標準プロファイル（JPGIS）第 2.1 版を用いて記述され，鉄道，路線，駅に関する情報を含む [5]．なお，Twitter のジオタグは WGS84 測地系を，鉄道時系列データは JGD2000 測地系であるが，差は数 cm〜m 程度なのでそのまま用いた．

3.4 実路線と近似直線の平方根平均二乗誤差

我々は，既に，実路線が存在する矩形領域で近似直線が抽出できたどうかを再現率を用いて評価し，再現率が低くなる原因について分析した [2]．ただし，単に近似直線の有無を調べているだけで，実路線とは異なる位置や方向である可能性もある．そこで，さらに実路線に本手法で抽出した近似直線がどの程度一致しているかを，実経路と推定した近似直線の距離と角度の誤差を用いて評価する．

まず，鉄道時系列データから JR 山手線の地点のリストを取り出して，本手法の矩形領域単位に分割し，矩形領域ごとに距離と角度の誤差を計算する．なお，実路線は存在するが測位点が存在しない領域がわずかに存在するが，この領域に関しては評価対象外とした．

距離に関しては，各地点と近似直線に垂直に引いた直線との交点までの距離 d を，ヒュベニの公式を用いて求める．

$$d = \sqrt{(d_y M)^2 + (d_x N \cos \mu_y)^2} \tag{3}$$

d_x，d_y は 2 地点間の経度・緯度の差，M は子午線曲率半径，N は卯酉線局率半径，μ_y は緯度の平均値である．なお，M や N を求める際に必要な赤道半径，扁平率の逆数は WGS84 測地系に従い 6,378,137m，298.257,223,563 とした [6]．

角度に関しては，少なくとも片方が矩形領域内に存在する隣接した 2 地点を通る直線と近似直線とがなす角度 r を求める．

最後に，各矩形領域の距離誤差 $RMSE_d$ と角度誤差 $RMSE_r$ を，平方根平均二乗誤差 ($RMSE$: Root Mean Square Error) を用いて計算する．

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N} e_i^2} \tag{4}$$

N はある矩形領域内の実路線の地点数または直線数，e_i は i 番目の距離 d または角度 r である．

3.5 実路線に対する近似直線の再現率の分析

抽出した近似直線の再現性を調べるために，誤差が閾値以下の場合に正しく抽出されたとして再現率を求めた．閾値は，測地系が異なる 2 地点間の距離を計算していること，本手法では近似直線では環状の山手線の曲線部で誤差が生じやすいこと，実路線データの測位間隔は均等でなく矩形領域によっては実際以上に悪い結果が生じやすいことを考慮して，誤差値の分布と近似直線の確からしさを調べて，50m と 30° とした．JR 山手線が通過する領域数，近似直線が抽出された領域数，$RMSE_d \leq$ 50m と $RMSE_r \leq 30°$ のどちらか片方または両方の条件が成り立つ場合の再現率を，表 1 に示す．$RMSE_d \leq 50\text{m} \wedge RMSE_r \leq 30°$ の場合でも，近似直線は 153 領域中 104 領域で正しく抽出され，再現率は 0.68 だった．

そこで，鉄道時系列データの地点と本手法の近似直線を Google Maps 上に可視化して調べると，再現性が悪かった矩形領域は全体に平均的に分散しているわけではなく，局所的に集中していることがわかった．特に再現性が悪かった JR 新宿駅周辺の可視化結果を図 3 に示す．

既に再現性が悪化する原因として，Twitter アクティブユーザ数が少ないこと，正確な位置取得ができない地下部分である

*1 https://dev.twitter.com/docs/api/streaming
*2 http://nlftp.mlit.go.jp/ksj/gml/datalist/KsjTmplt-N05.html

表 1: JR 山手線の各条件下の路線数と再現率

	実路線	近似直線の有無	$RMSE_d \leq 50\text{m}$	$RMSE_r \leq 30°$	$RMSE_d \leq 50\text{m} \wedge RMSE_r \leq 30°$
路線数	153	142	117	109	104
再現率	–	0.928	0.765	0.712	0.680

図 3: 再現性が悪い場所の可視化結果（JR 新宿駅周辺）

ことを示した [2] が，主要駅周辺にも新たな再現性が悪い原因を発見した．

一つは，主要駅周辺には，職場，店，レストランなどの非常に多くの訪問地点が分散していることである．本手法は矩形領域内に明らかにツイート密度が高い連続部分が存在するという仮定に基づくので，ツイート密度の変化が小さい領域ではうまく抽出できない．図 3 を見ると，新宿駅およびその北側の矩形領域では，近似直線の角度に大きな誤差が生じていることがわかる．

もう一つは，主要駅は，複数の路線の乗り換え駅だったり，主要道路がすぐ近くを通過していることが多いことである．例えば，新宿駅は鉄道は山手線・埼京線に加えて中央本線・総武線，小田急線，京王線などが，主要道路としては青梅街道，甲州街道，首都高速などが交差している．現時点の手法の制約から，このような場合でも一つの矩形領域から 1 本の近似直線しか抽出しない．図 3 を見ると，新宿駅南側で山手線沿いではなく首都高速または京王線方面沿いに近似直線が抽出されていることがわかる．また，交差部分の路線形状が複雑な場合や角度差が小さい場合には，先ほどと同様にツイート密度の変化が小さくなり，うまく抽出できない．図 3 の上部の西武新宿駅の北側は，山手線と中央本線が分岐していく場所であり，抽出された近似直線の角度に大きな誤差が生じている．

4. おわりに

本稿では，交通路の抽出を行う本手法に対して，どの程度正確に近似直線の抽出を行えているのかの評価を行った．実路線の地点が存在する領域数 153 中 104 領域数 (68.0%) 正確に再現されていた．正確な再現が不可な領域はまず，人口密度が高い主要駅を含む領域で，周辺施設の充実や，路線の乗換駅のため，ノイズが多く混入することが原因であると考えられる．また，複数路線存在する矩形領域では，別路線の抽出や，移動ツイートが入り交じり特定路線の抽出を行えかったことが原因として考えられる．

今後の課題としては，今回の分析結果を考慮し，現在の移動ツイートからさらにノイズを除去することや，同一経路としてまとめたグループの欠損部分に近似直線を補完することで複数路線抽出を行うことである．さらに，ベジエ曲線などを用いたなめらかな交通路抽出により誤差の縮小が可能であると考える．

謝辞

本研究は JSPS 科研費 26330345 の助成を受けた．

参考文献

[1] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes twitter users:real-time event detection by social sensors. In *Proceedings of the 19th International Conference on World Wide Web (WWW '10)*, pp. 851–860, 2010.

[2] 谷直樹, 風間一洋, 榊剛史, 吉田光男. ジオタグ付きツイートを用いた交通路の抽出. *DEIM 2015 F7-4*, 2015.

[3] Yasunori Hada, Takeyasu Suzuki, and Itsuki Noda. Utilization of probe vehicle information in disasters in japan. In *Proceeding of the 15th World Conference on Earthquake Engineering (WCEE2012)*, 2012.

[4] Richard O. Duda and Peter E. Hart. Use of the hough transformation to detect lines and curves in pictures. *Communications of the ACM*, Vol. 15, No. 1, pp. 11–15, 1972.

[5] 国土交通省国土地理院. 地理情報標準プロファイル (JPGIS) Ver. 2.1, 2009.

[6] 吉田聡, 古屋貴司, 稲垣景子. 図解! ArcGIS 10 Part1 身近な事例で学ぼう. 古今書院, 2012.