情報処理学会論文誌　データベース　Vol.10　No.2　31–41 (June 2017)

ジオタグ付きツイートを用いた交通路の抽出法

谷 直樹[1,a)]　風間 一洋[2,b)]　榊 剛史[3,c)]　吉田 光男[4,d)]　斉藤 和巳[5,e)]

受付日 2016年12月10日, 採録日 2017年4月4日

概要：Twitter は現実世界の状況を把握するソーシャルセンサとして活用されており，位置情報が付加されているジオタグ付きツイートを用いた人間の移動や観光地情報の分析がさかんに行われている．本論文では，ジオタグ付きツイートを内容や移動速度，移動距離などの条件に基づいて選別・集積して得られる位置情報から，それらの発言者たちが共通で利用している交通路を抽出する方法を提案する．実際には，まずツイート投稿中またはその前後にユーザが移動したと推定されるツイート群を抽出し，対象地域を細分化した矩形領域内のツイートを Hough 変換して，交通路の断片である近似線分群を求める．次に，連続していると推定される近似線分をグループ化して，3 次スプライン曲線で近似・補間することで，連続した交通路として抽出する．実際に，JR 山手線周辺の領域の抽出結果を可視化して，提案手法の特徴を分析する．また，特に鉄道路線に着目して，国土数値情報鉄道時系列データと比較することで，提案手法を評価する．さらに，動的に生成される経路抽出の応用例として，桜並木に沿って移動する花見客の経路を分析する．

キーワード：Twitter，ジオタグ，経路抽出，Hough 変換，3 次スプライン補間

Traffic Route Extraction Method from Geotagged Tweets

Naoki Tani[1,a)]　Kazuhiro Kazama[2,b)]　Takeshi Sakaki[3,c)]　Mitsuo Yoshida[4,d)]
Kazumi Saito[5,e)]

Received: December 10, 2016, Accepted: April 4, 2017

Abstract: Twitter is used as a social sensor to grasp the situation of the real world, especially it has been an actively conducted research using a geotag, such as analysis of human movement or tourism attractions information. In this paper, we propose a method to extract traffic routes of public transports that many users used. Specifically, we obtain an approximate straight line, which seems to be a fragment of a traffic route, in each rectangle area by Hough transform of positions that are obtained from tweets during moving, and we make groups of approximate straight lines, and curve by spline interpolation. Actually, it is visualized extracted traffic routes around the JR Yamanote Line from geotagged tweets. Furthermore, we show the effectiveness of methods by evaluating the recall/precision of an approximate straight line extraction and the degree of grouping. Additionally, we evaluating the performance of our method by comparison with National Land Numerical Information Railway Data. Furthermore, we analyze the route of people enjoying the cherry blossoms by our method in an application example.

Keywords: Twitter, geotagging, route extraction, Hough transform, cubic spline interpolation

[1] 和歌山大学大学院システム工学研究科
Graduate School of Systems Engineering, Wakayama University, Wakayama 640–8510, Japan
[2] 和歌山大学システム工学部
Faculty of Systems, Wakayama University, Wakayama 640–8510, Japan
[3] 株式会社ホットリンク
Hotto Link Inc., Chiyoda, Tokyo 102–0071, Japan
[4] 豊橋技術科学大学情報・知能工学系
Department of Computer Science and Engineering, Toyohashi University of Technology, Toyohashi, Aichi 441–8580, Japan
[5] 静岡県立大学経営情報学部
School of Management and Information, University of Shizuoka, Shizuoka 422–8526, Japan
[a)] s161032@center.wakayama-u.ac.jp
[b)] kazama@sys.wakayama-u.ac.jp
[c)] t.sakaki@hottolink.co.jp
[d)] yoshida@cs.tut.ac.jp
[e)] k-saito@u-shizuoka-ken.ac.jp

1. はじめに

スマートフォンやタブレットなどのモバイル端末の普及にともない，Twitter に代表されるソーシャルメディアの利用が急増し，様々な情報を収集・交換できる重要な情報インフラとして注目を浴びている．特にユーザの発言位置が付加されるジオタグ付きツイートを用いて，ツイートの内容とユーザの位置を組み合わせて分析することで，たとえば観光地情報の分析や，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として活用できることが知られている [1], [2].

本論文では，内容や移動速度，移動距離などの条件に基づいて選別・集積して得られるジオタグ付きツイートの位置情報から，それらの発言者たちが共通で利用している経路を発見する手法を提案する．位置情報の選別手法を目的に合わせて変更すれば，提案手法では，特に複数のユーザが同じ目的で行動をともにした経路部分が抽出されることから，東日本大震災時に提供された「通れたマップ」[3] のように震災直後に車が通行可能な経路や，桜並木がある場所，定番の観光ルートなど短期間であっても比較的多くのユーザが密集する状況下で動的に生成される経路を推定できると思われる．

本手法を実現するために，まずツイート投稿中またはその前後にユーザが移動したと推定されるツイート群を抽出し，対象地域を細分化した矩形領域内のツイートを Hough 変換して，交通路の断片である近似線分を求める．次に，連続していると推定される近似線分をグループ化して，3 次スプライン曲線で近似・補間することで，連続した交通路として抽出する．これによって，ツイートは散発的にしか行われない，位置取得対象が移動手段ではなくユーザである，単一ユーザの位置情報だけでは経路を忠実に再現できず実際にどのような移動手段を利用したかも正確に判定できない，といった既存の経路検出手法では適用できない問題を解決する．実際に，JR 山手線周辺の領域の抽出結果を可視化して，提案手法の特徴を分析する．さらに，特に鉄道路線に着目して，国土数値情報鉄道時系列データと比較することで，提案手法の有効性を示す．

2. 関連研究

移動手段に取り付けた GPS を用いて継続的に短い間隔で取得した位置情報から，移動手段の経路を抽出する研究が存在する．Hada らは，多数のプローブカー（GPS と通信機能を搭載した自動車）の位置を収集して地図上に可視化することで，震災時に通行可能な経路を抽出した [4]．山田らは，ユーザの携帯 GPS を用いて 5 分間隔で収集した位置情報履歴を利用して，現在の経路と時間的・空間的に同一経路と判定された過去の移動経路の通過確率を求めて，訪問地・経由地を推定した [5]．Wang らは，車に設置した GPS の位置データから作成した，運転者，地図と照合して求めた道路ネットワーク上の道路区間，タイムスロットから作られた 3 次元のテンソルを分解して，2 地点間の移動経路と移動時間を推定した [6]．Wei らは，Foursquare のチェックインデータや北京におけるタクシーの移動データから抽出した位置情報を，格子状に分割したセルに割り当てて，それらが結合する領域を求めた後に領域内・間のエッジを推定しておき，与えられたクエリに対して時間的，空間的に類似している上位 k 件の経路を推定するフレームワーク RICK を提案した [7]．Krumm は，目的地推定のために，GPS ロガーを設置した複数の自動車の軌跡を統合して地図と照合することで，精度が高い経路を求める手法を提案している [8]．

また，ユーザが携帯する GPS を用いて継続的に短い時間間隔で取得した位置情報から得られるユーザの移動軌跡から，同一移動手段（例：徒歩，車，電車など）を用いた区間を推定する研究が行われている．Reddy らは，ユーザの GPS 軌跡データから計算される速度・加速度などの特徴から，移動手段を判定する分類器を構築した [9]．Stenneth らは，ユーザの位置を地図上のバス停や駅などの位置と照合することで，速度だけでは判別が困難な電車・バス・自動車などに対しても移動手段を推定できることを示した [10]．Zheng らは，速度・加速度・移動距離に加えて，大きく変化している速度・進行方向の GPS 測位点や停止している GPS 測位点の割合を考慮し，移動形態の推定を行った [11]．また遠藤らは，GPS の移動軌跡からいったん軌跡画像を生成し，その特徴を深層学習を用いて学習することで，移動手段を推定した [12].

他に，位置情報ではなくテキスト情報を用いた研究として，石野らは，被災時のユーザ行動経路情報を含むツイートに含まれる移動元，移動先，移動手段を表す単語を，CRF 法を用いてタグ付けすることで，ユーザの移動経路を抽出する手法を提案した [13]．移動軌跡ではないが，大森らは，Flickr の写真に付与されるタグは撮影位置の特徴を表していると仮定して，ユーザが「beach」とタグ付けた写真の位置だけを抽出して，連続した海岸線を推定する手法を提案した [14].

以上で述べた既存手法では，移動経路上の位置を連続した経路として取得できているか，テキストやタグなどの情報を用いて抽出対象に直接関係する位置だけに絞り込めることが前提である．本手法では，ツイート時の位置しか利用できないために測定間隔は長く，測定点数は少ないため，位置情報が連続した軌跡として得られないうえに，無関係な位置情報が多数存在するような条件下で，複数ユーザの位置情報を統合することで密度が高い連続部分として創出される経路の抽出を試みる点が異なる．

3. 提案手法

3.1 利用可能なデータの問題点と手法の概要

GPS を使って移動手段やユーザの位置を短い間隔で継続的に取得できる場合と異なり，ジオタグ付きツイートから交通路を抽出する場合には，以下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：ツイート時にしか位置を取得できず，そのタイミングはユーザの行動に依存する．通常は取得間隔は長く不定であり，単一ユーザだけでは移動軌跡を再現できない．

(2) 位置取得対象の問題：移動手段ではなくユーザの位置なので，すべてが移動手段の経路上にあるとは限らない．

(3) 移動手段の利用判定の問題：前記の理由から，ユーザの移動手段の利用を判定する必要があるが，連続するツイートの間の移動の有無は推定できても，ツイート時の移動の有無は分からない．

本論文では，以下のような複数ユーザの位置情報を統合することにより，密度が高い連続部分として創出される経路を抽出する手法を提案する．

(1) 移動ツイート集合の作成：ツイートの投稿時刻と位置から，少なくとも発言前後に移動があったツイート集合を求める．

(2) 近似線分の抽出：地理空間上を矩形領域に分割し，各領域のツイート密度が高い連続部分を Hough 変換により経路を示唆する近似線分として抽出する．

(3) 同一経路の近似線分のグループ化：まず局所的な連続性に基づいて隣接する近似線分を接続し，さらに求めた近似線分群を広域的な連続性に基づいて接続しなおすことで，同一経路と推定される近似線分グループを取得する．

(4) 3 次スプライン曲線による補間：各近似線分グループを 3 次スプライン曲線で補完して，連続した経路として抽出する．

3.2 移動ツイート集合の作成

自動車，バス，電車，新幹線などの交通手段では，運転者でない限りはツイートできるが，移動手段の利用判定の問題から，ツイート時に何らかの交通手段を利用していたかは分からない．そこで，個々のユーザごとに抽出される連続する 2 つのツイートの投稿位置と時間から求めた移動速度から，ツイート前後で移動したと推定されたツイート集合を作成する．

移動判定の概念図を，**図 1** に示す．まず，ある特定のユーザのツイートの時系列 $\{TW_0, TW_1, TW_2, TW_3, TW_4, TW_5, TW_6\}$ が与えられ，ツイート TW_i を地点 p_i で時刻 t_i に投稿したと考える．ツイート地点 $p_0 \sim p_6$ のうち，p_0 は出発地（自宅），p_6 は目的地（レストラン），$p_1 \sim p_5$ は

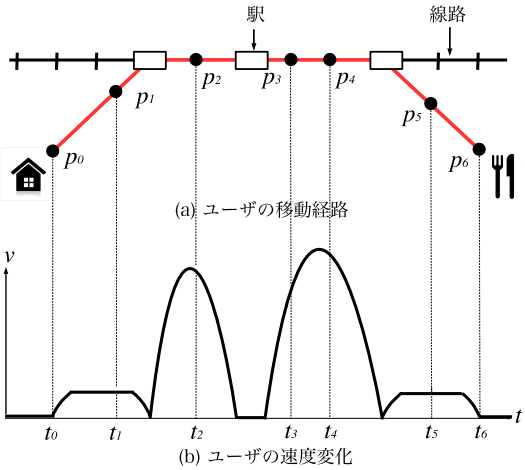


図 1　ツイートの移動判定の概念図

Fig. 1　Conceptual diagram of movement determination using tweets.

移動中のツイート地点であり，そのうち p_1，p_5 は徒歩中，p_2，p_3，p_4 は電車に乗車中である．

本来は，ユーザの移動速度 v_i は，ツイート時刻 $t_0 \sim t_6$ に対して，図 1 に示すように連続的に変化すると考えられるが，実際には速度 v_i は分からない．そこで，代わりに地点 p_i，p_{i+1} の間の停止時間を含めて計算した速度の平均値である表定速度 $\bar{v}_{i,i+1}$ を用いる．

$$\bar{v}_{i,i+1} = \frac{d(p_i, p_{i+1})}{t_{i+1} - t_i} \tag{1}$$

$d(p_i, p_{i+1})$ は，p_i，p_{i+1} 間の直線距離を求める関数である．この距離は，本来ならばユーザの正確な移動経路から求めることが望ましいが，散発的で少数のツイートからでは経路の把握が困難であり，直線距離ではノイズの混入が考えられる．

そこで，精度を重視したツイートの抽出を目的として，p_i が次の条件を満たす場合にユーザが移動中であると考え，ツイート TW_i を**移動ツイート**と呼ぶ．

$$\bar{v}_{min} \le \bar{v}_{i-1,i} \le \bar{v}_{max} \wedge \bar{v}_{min} \le \bar{v}_{i,i+1} \le \bar{v}_{max} \tag{2}$$

$\bar{v}_{min}$ は対象とする移動手段の表定速度の下限値，$\bar{v}_{max}$ は上限値とする．ここで $\bar{v}_{min}$，$\bar{v}_{max}$ を用いる理由は，位置系列が離散的すぎて，カルマンフィルタなどを用いた異常値除去が困難だからである．

たとえば，電車を想定した場合は，$\bar{v}_{1,2} \ge \bar{v}_{min}$ でも，$\bar{v}_{0,1} < \bar{v}_{min}$ なので TW_1 は移動ツイートではないが，$\bar{v}_{2,3} \ge \bar{v}_{min}$ なので TW_2 は移動ツイートとなる．ただし，$\bar{v}_{1,2} < \bar{v}_{min}$ なら，TW_2 は移動ツイートと判定されない，すなわち，電車乗車中の最初と最後のツイートが除外される可能性があるが，本論文のようにユーザ数が多い場合には，精度を重視して確実に移動中のツイートだけを抽出する．

なお，この処理は，ある条件に合致するツイート群を抽出

する処理として一般化でき，$\bar{v}_{min}$，$\bar{v}_{max}$ を調整したり，新たな条件を追加したりすることができる．たとえば，$\bar{v}_{min}$ を高めの値に設定して新幹線に絞り込んだり，ツイートの内容や移動距離などの条件を考慮して，移動の目的を限定したり，抽出精度を高めたりすることができる．

3.3　近似線分の抽出

対象地理空間を分割した各矩形領域に対して，**Hough 変換** [15] を用いて移動ツイート群の特に密度が高い連続部分を経路の部分的な近似線分として抽出する．

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して，エッジ群を通る直線を抽出する手法である．エッジの始点・終点の x，y 座標と，原点からの距離 ρ，角度 θ の間には，次の式が成り立つ．

$$\rho = x\cos\theta + y\sin\theta \tag{3}$$

ただし，ツイート間隔は比較的長いことから，同一ユーザから得られるツイート数は少なく，移動経路を再現するのは困難である．そこで，同一ユーザに限定せずに，単に近接する 2 つの移動ツイートの位置をつないだエッジ集合を入力データとする．ユーザの経路上にない組合せでもエッジが作られるので，抽出されたエッジ集合を対象とする通常の Hough 変換の場合よりも大量のノイズが混在するが，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を除去する．

各矩形領域の近似線分の抽出アルゴリズムを，以下に示す．

(1) 矩形領域に存在するユーザが少ない場合は，目視で観察したところ，特定ユーザの自宅や職場であったので，経路と関係がない可能性を考慮し，閾値 T_u を下回る場合は終了する．
(2) 距離が閾値 T_d 以下の 2 つのツイートの位置を通過するエッジ群を求める．
(3) エッジ群を Hough 変換し，Hough 空間上の距離・角度で分割した各領域の頻度を集計する．
(4) G 検定で Hough 空間上の偏りがないと判定された場合は，終了する．
(5) Hough 空間上でエッジ数が最大の領域の距離，角度の平均値を求め，線分として逆写像する．

G 検定は，帰無仮説に対して実際に観測された理論分布との異なりを検定する統計学的検定法であり，カイ二乗検定のように対数尤度を近似せずに直接用いることから，標本サイズの影響がないという利点を持つ．矩形領域内の分布の偏りの有無は，以下の手順で判定する．

(1)「Hough 空間上の分布は均一である」という帰無仮説をたてる．
(2) Hough 空間上の各区間のエッジ出現頻度を O_i，帰無仮説で期待される出現頻度であるエッジの平均値を E_i として，G 値を求める．

$$G = 2\sum_i O_i \times ln(\frac{O_i}{E_i}) \tag{4}$$

(3) G 値が有意水準 α 以下なら，帰無仮説を棄却する．

3.4　同一経路の近似線分のグループ化

次に，抽出した国道，路線，高速道路のような長い経路を再現するために，同一経路上にあると推定される近似線分に対して，2 段階のグループ化を行う．

3.4.1　近似線分の局所的接続

まず，局所的な連続性に基づいて近似線分を接続するアルゴリズムを以下に示す．なお，近似線分が**図 2** のように配置されているとし，l と l' の中点を結ぶ直線を引いた場合の l の角度を α，l' の角度を β，l と l' の中線の角度を γ とする．また，$diff_\theta(\theta,\theta')$ は，角度 θ，θ' の角度差 θ'' $(0 \le \theta'' < 360)$ を，$diff_d(d,d')$ は距離差 d'' 求める関数であり，T_{θ_1}，T_{θ_2} は角度，T_d は距離の閾値とする．

(1) 矩形領域 (x,y) の近似線分 l の周囲の $(x-1,y-1)$，$(x+1,y-1)$，$(x-1,y+1)$，$(x+1,y+1)$ で囲まれる 8 領域から，次の条件を満たす近似線分集合を求める．

$$\begin{cases} diff_\theta(\alpha,\gamma) \le T_{\theta_1} \wedge diff_d(\beta,\gamma) \le T_{\theta_1} \\ diff_d(d_1,d_2) \le T_d \end{cases} \tag{5}$$

(2) 近似線分が存在しない場合は，トンネルのように測位できないために一部が断続している可能性を考慮して，さらにその外側の $(x-2,y-2)$，$(x+2,y-2)$，$(x-2,y+2)$，$(x+2,y+2)$ で囲まれる 16 領域に拡大する．
(3) 近似線分集合から，近似線分 l' を取り出す．存在しない場合は，終了する．
(4) l の端点 e_1 に最も近い近似線分 l' に対して，l に対する l' の連続度 $C(l,l')$ と l' に対する l の連続度 $C(l',l)$

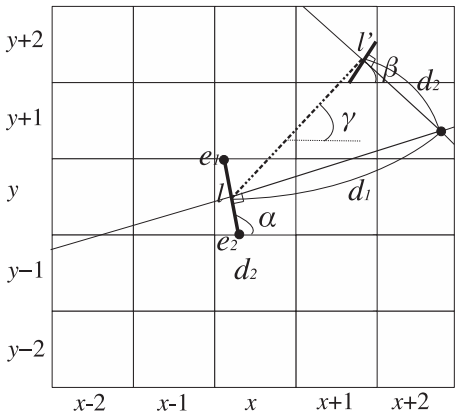


図 2　近似線分同一経路集合を構築する例

Fig. 2　Example of constructing an approximate line set for the same route.

が最も高い場合に，l' を接続先として，グループに追加する．

(5) e_2 に対しても，(4)〜(5) と同様の処理を行う．

(6) (3) に戻る

なお，l と l' の連続度 $C(l,l')$ は，距離差 $diff_d(d_1,d_2)$ と，角度差 $diff_\theta(\alpha,\beta)$ として，以下のように求める．

$$C(l,l') = \frac{(1-\frac{diff_\theta(\alpha,\beta)}{T_{\theta_2}})+(1-\frac{diff_d(d_1,d_2)}{T_d})}{2} \qquad (6)$$

3.4.2　近似線分群の広域的接続

近似線分の局所的接続では，経路上にない近似線分を接続するなどの誤接続が発生するために，大域的な観点から近似線分がより適切に接続するように，以下のアルゴリズムで再グループ化する．

(1) グループ間の接続可能性行列を作成する．なお，グループ G_i と G_j が式 (5) を満たす場合には i 行 j 列の値を 1，そうでなければ 0 とする．

(2) 接続可能性を持たない端となる近似線分を持つグループ G_i を探索する．

(3) グループ G_i から接続可能なすべてのグループ系列に対して，連続度 $C(l,l')$ の総和を計算する．

(4) 連続度 $C(l,l')$ の総和が最大のグループを経路集合に格納する．

(5) 最大グループを形成するグループ系列の各グループから，最大グループに含まれる近似線分を取り出す．

(6) 最大グループのサイズが T_g 未満なら終了する．それ以外なら，新たなグループ間の接続可能性行列を作成し，(2) に戻る．

ただし，サイズが非常に小さいグループは，国道，路線，高速道路など経路は連続して存在することを考えると，ノイズである可能性が高く，閾値 T_g より小さいグループは除外する．

3.5　3 次スプライン曲線による補間

再抽出によりグループのサイズは大きく経路の連続性は高くなるが，近似線分にギャップが存在したり，曲線部の再現性が低くなったりするという問題がある．そこで，同一経路を表すと思われるグループに属する近似線分の中点集合を求めて，**3 次スプライン曲線**を用いて補間する．

この補間により，近似線分のギャップが解消され，さらに曲線部の再現性やノイズによる揺らぎが解消される．

4.　評価

本提案手法により，発言者たちが共通で利用している交通路を抽出できているかを確認するため，まず関東の JR 山手線周辺領域を可視化するとともに，各路線に対してどの程度正確に抽出できているかを評価する．

また，このような恒常的に利用可能な自明な経路以外に，比較的多くのユーザが密集する状況下で動的に生成される経路を抽出可能かを確認するため，どの程度のツイート数があれば経路を抽出可能であるかを実験し，動的に生成される経路として，花見の時期の桜並木の抽出結果を分析する．

4.1　ツイートデータセット

Twitter Streaming API [*1] を用いて，ジオタグ付きツイートだけを収集した JSON 形式のツイートデータセットから，2012 年 5 月 1 日〜2014 年 4 月 30 日の 2 年分を評価に使用した．ただし，ジオタグを付加してツイートするボットが存在することから，人間のつぶやきだけを分析対象にするために，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）に「BOT」などの文字列が含まれているアカウントをボットと見なして除去した．使用したデータセットに含まれるツイート数は 177,069,705，ユーザ数は 2,409,050 であり，実験対象は，十分なツイートデータが得られることや複数の路線・道路が存在するといった点から関東の JR 山手線周辺区域 1,150 km^2 とし，そこからは 14,400,828 ツイート，265,977 ユーザ，さらに移動ツイートに関しては 52,033 のユーザから 261,911 ツイート得られた．

4.2　パラメータの設定

多くの人が利用する公共交通機関の抽出を想定して，**表 1** のように各種パラメータを設定した．移動ツイート集合作成の対象速度は，下限値 $\bar{v}_{min}$ を各駅停車の速度の 18 km/h，上限値 $\bar{v}_{max}$ を新幹線の最高速度の 285 km/h とした．交通路の近似線分の抽出では，Hough 空間は細かく分割した際ツイートが集中しないため，距離軸は 250 m 矩形領域の対角線の 1/3 の 117 m，また国道や路線などの横幅を考慮して 10° の領域に分割した．さらに，ユーザ数の閾値は 1 では自宅や職場，2 ではそれらに加えて位置測位誤差が考えられるため，T_u を 3 とした．同一経路と推定される近似線分のグループ化では，T_{θ_2} の値を角度差の最大値の 90° とした．

その他の閾値は，最も性能が良かった値を用いた．

4.3　抽出結果の可視化

交通路の可視化結果を，**図 3** に示す．全般的に，電車や新幹線，高速道路は比較的忠実に抽出できた．ただし，電車の場合は，東京駅などの主要乗り換え駅周辺は複数路線が乗り入れているうえに，形状が複雑なことからうまく再現できなかった．新幹線の場合は，直線的な区間において一部欠落がみられた．高速道路の場合は新幹線より欠落区間が多く，一般道は一部が断片的に抽出されただけであっ

[*1] https://dev.twitter.com/docs/api/streaming

表 1　本論文で用いたパラメータ
Table 1　Parameters used in this paper.

対象速度	矩形領域			Hough 空間		G 検定	グループ化			
$\bar{v}_{min}$～$\bar{v}_{max}$	面積	ユーザ T_u	ツイート距離 T_d	距離 ρ	角度 θ	α	T_{θ_1}	T_{θ_2}	T_d	T_g
18～285 km/h	250 m	3	176 m	117 m	10°	0.01	45°	90°	100 m	4

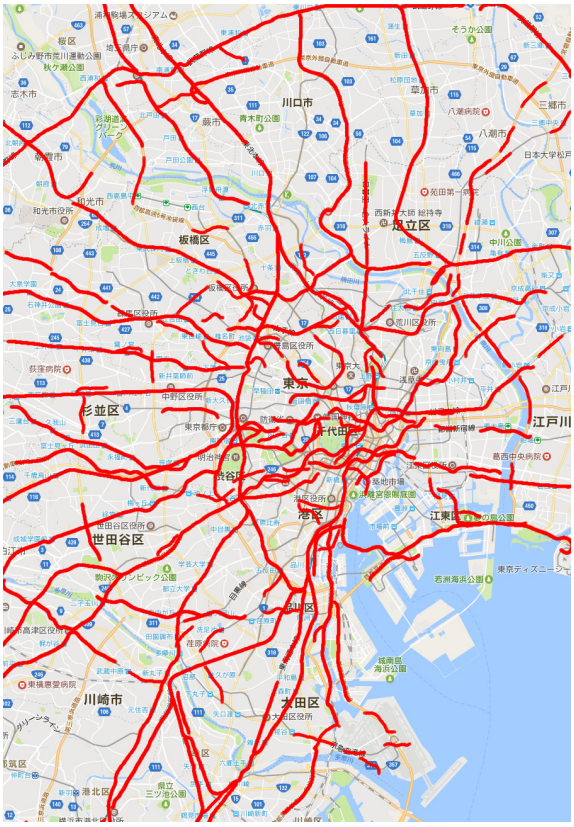


図 3　JR 山手線周辺区域の可視化結果
Fig. 3　Visualization result in the surrounding area of JR Yamanote line.

た．これは運転中にツイートできるのは同乗者に限られるために，経路上のツイート数が少なくなるからと思われる．また，郊外部は，他の路線や道路の影響を受けにくく，直線状であることから，高精度に抽出が行えていることが確認できた．

そこで，以降は比較的良好に抽出された鉄道路線に対して評価する．

4.4　評価用路線データ

鉄道路線の抽出性能を評価するために，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の鉄道時系列データ[*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

[*2] http://nlftp.mlit.go.jp/ksj/gml/datalist/KsjTmplt-N05.html

表 2　鉄道時系列パッケージの代表的なタグ
Table 2　Typical tags in train time-series package.

クラス	属性・関連役割	タグ名	型
鉄道	事業者種別	int	事業者種別コード
	路線名	lin	string
	運営会社	opc	string
路線	路線	loc	CurvePropertyType
駅	地点	loc	PointPropertyType
	駅名	stn	string

され，鉄道，路線，駅に関する情報を含んでいる [16]．

代表的なタグを，**表 2** に示す [17]．型 CurvePropertyType は複数地点の緯度・経度のリストであり，型 PointPropertyType は 1 点の緯度・経度である．なお，Twitter のジオタグは WGS84 測地系を用いているが，鉄道時系列データは JGD2000 測地系を用いているので，位置の誤差が生じることが考えられるが，実際の差は数 cm から数 m 程度であることから，そのまま使用した．

さらに，opc タグと lin タグを手がかりに路線を判定し，さらに loc タグから路線の緯度・経度のリストを取得した．なお，lin タグの路線名は一般的に使われる路線名と 1 対 1 対応しているわけではなく，日比谷線のように「2 号線日比谷線」となっていたり，山手線のように，区画ごとに「山手線」，「東海道線」，「東北線」と複数に分割されていた．そこで，実際の路線名と鉄道時系列データは，人手で対応付けを行った．

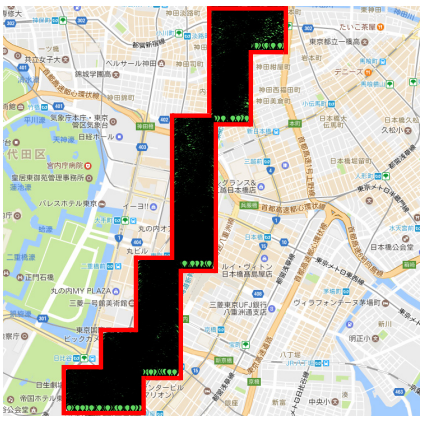
4.5　移動ツイート集合の抽出性能の評価

移動ツイート集合に含まれるツイートを確認するため，対象領域から抽出可能な総ツイートと，移動ツイート集合を比較する．

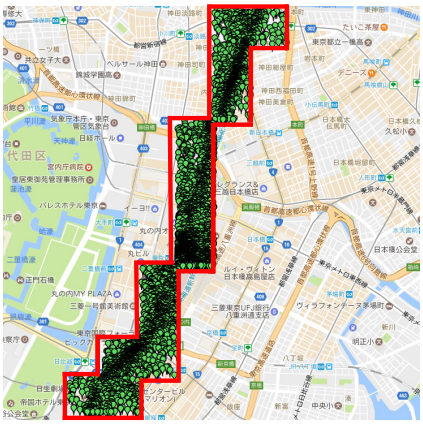
評価用路線データから JR 山手線上の地点リストを抽出し，その地点が含まれる東京駅周辺の矩形 14 領域から抽出可能な総ツイートと，移動ツイート集合を可視化した結果を**図 4** に示す．なお，ツイート位置は緑色のマーカで，また対象とする 14 領域のマーカを，赤色の線で囲んでいる．

まず，総ツイートの可視化結果を示す図 4 (a) では，どの領域もマーカのエッジにより，全体的に黒く可視化されている．これは，多数のマーカが密集しているためであり，経路以外にも様々な地点でツイートが取得されていることが分かる．

一方，移動ツイート集合の可視化結果を示す図 4 (b) は，中央の東京駅周辺では濃淡はそれほど強く見られないが，



(a) ツイート集合の可視化結果



(b) 移動ツイート集合の可視化結果

図 4　東京駅周辺の JR 山手線上のツイート位置の可視化結果
Fig. 4　Visualization result of tweet positions on JR Yamanote line surrounding Tokyo station.

上部と下部では，ある地点においては黒く可視化されており，ツイートの疎密が明確に見られる．このようにツイートが密である地点を確認したところ，山手線や東海道新幹線の路線上であり，疎であるところの多くは，路線上でない地点であった．つまり，路線上で多くのツイートが取得され，それ以外では少なく，総ツイートと比べて，目的としている移動時のツイートだけを多数抽出できていることが分かる．

4.6　距離誤差の評価

本手法で抽出した交通路が実路線をどの程度忠実に再現しているかを知るために，実路線と抽出結果の距離 d_i の平方根平均二乗誤差（$RMSE$: Root Mean Square Error）を次のように計算した．

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N} d_i^2} \tag{7}$$

N は矩形領域内の測定数，d_i は計測地点 i における実路線と抽出結果の距離である．ただし，鉄道時系列データにおける計測間隔は統一されておらず，実路線が通過する矩型領域内の計測地点が存在しないこともあったので，擬似的に実路線の計測地点の間を直線で結び，より細かい間隔でその直線と抽出結果の間の距離を調べた．

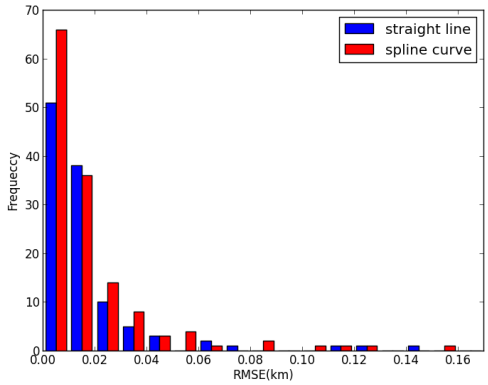


図 5　距離誤差 $RMSE$ のヒストグラム
Fig. 5　Histgram of RMSE.

距離 d はヒュベニの公式を用いて次のように求める．

$$d = \sqrt{(d_y M)^2 + (d_x N \cos \mu_y)^2} \tag{8}$$

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 6,378,137 m，298.257,223,563 とした [18]．

山手線の場合の各矩型領域の RMSE 値のヒストグラムを**図 5** に示す．横軸は RMSE（km），縦軸は頻度を表す．また，赤は近似線分の場合，青はスプライン曲線の場合である．この結果から，スプライン補完した場合には，誤差が一番小さい 0.00～0.01 の区間で 51 から 66，比較的誤差が小さいと考えられる 0.00～0.05 の区間で頻度が 20 向上した．実際に山手線路線の抽出結果を確認すると，池袋駅～大塚駅間の曲線部や，渋谷駅を含む複数路線混入している領域で曲線が再現されており，また大塚駅～巣鴨駅間の曲線部で誤差が減少していたなどスプライン補間による有効性が確認できる．

4.7　交通路の再現性の評価

本手法による交通路の再現性は，交通路網の構造と，対象領域の Twitter アクティブユーザ・移動ツイート数，通信と位置取得に用いる GPS 衛星・携帯基地局・Wi-Fi アクセスポイントの電波強度と補足数などに大きく左右される．そこで，本手法で交通路がどの程度再現できているかを各路線の Precision と Recall で評価した．

まず，抽出結果に対して人手で路線タグを付与し，Precision と Recall を次のように計算した．

$$Precision = \frac{K}{D} \tag{9}$$

$$Recall = \frac{K}{C} \tag{10}$$

C は評価用路線データにある路線の測定地点が存在する矩

表 3　評価用路線データと F 尺度
Table 3　Railway data used in evaluation and F-measures.

	路線名	平均乗降客数	Precision	Recall	F 尺度
地上路線	山手線	447,640	0.547	0.830	0.659
	中央線	47,527	0.534	0.775	0.631
	東横線	106,149	0.333	0.840	0.474
	小田原線	66,543	0.255	0.737	0.378
	京王線	53,196	0.355	0.724	0.476
	目黒線	49,854	0.040	0.480	0.073
地下路線	日比谷線	63,055	-	0.151	-
	千代田線	74,466	-	0.133	-
	銀座線	56,473	-	0.044	-
	有楽町線	40,000	-	0.100	-

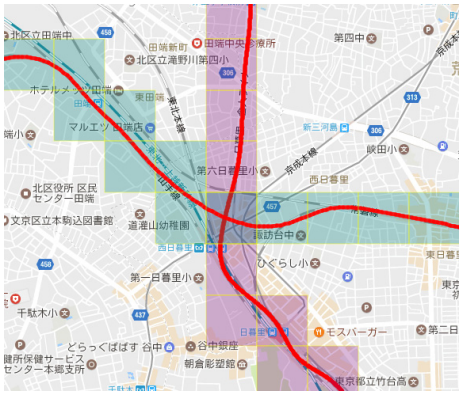


図 6　近似線分の誤接続が行われた箇所
Fig. 6　Area where approximate lines were misconnected.

形領域数，D はその路線タグが付与された抽出結果が属する矩形領域数，K は距離誤差 RMSE が閾値以内のその路線タグが付与された抽出結果が存在する矩形領域数である．RMSE の閾値は，明らかに再現困難な箇所を明確にしたいことから 50 m とした．

評価対象とした路線名，1 日あたりの 1 駅の平均乗降客数，Precion，Recall，F 尺度を，**表 3** に示す．

Recall に注目すると，地上路線は全体的に高い．この理由としては，今回評価対象とした JR 山手線周辺区域は人口 1000 万人以上であることから，ジオタグ付きでツイートするユーザ数が十分得られること，昼間の交通渋滞と電車網の普及から電車の利用率が高いこと，東京都心部の地価や家賃の高さから郊外部と都心部を往復するユーザが多いことなどが考えられる．一方，Recall が低下する原因としては，たとえば**図 6** に示す赤色で示される 2 つの曲線を見ると，山手線と常磐線，日暮里・舎人ライナー上の近似線分から構成されていることが確認できる．本来，1 路線 1 曲線で構成されることが望ましいが，このように誤ってグループ化された場合，スプライン補間の際に経路から線分がそれるため誤差が大きくなる．

また，地下路線は全体的に低い．この違いが出てくる理由は，地上では GPS 衛星の捕捉が可能で，さらに複数の

図 7　複数の路線の近似線分が誤ってグループ化された箇所
Fig. 7　Area where multiple lines were misgrouped.

Wi-Fi 基地局位置取得が可能であるのに対して，地下路線では位置取得に何らかの制約や問題があるからだと考えられる．地下鉄路線で再現されている箇所を見ても地上路線により再現されていると思われる．

また Precision は全体的に低い値を示した．Precision が低下した原因を分析するために，東横線の実路線の地点と路線タグが付与された曲線の可視化結果を**図 7** に示す．緑マーカが東横線の正解地点リスト，黄色の領域上の赤の曲線が路線タグが付与された曲線である．この曲線は，東横線以外に山手線と常磐線，また京成本線から構成される曲線であるため，このように路線ごとにセグメントとして分割されていない曲線では，その他の路線上の領域が加算されるため，Precision は低くなる．

4.8　ツイート数の変動に基づく抽出性能の評価

今までは，ツイート頻度が著しく低い区間でも抽出できるように，なるべく長期間のツイートを用いて評価を行った．本手法では，通行可能経路が不明になるような災害時や，一時的に道路を封鎖して行うイベントなど比較的短期間で多くのユーザが密集するような状況で，動的に作られる経路の抽出への応用も意図しており，このような場合も比較的ツイート数が得られると考えられるが，どの程度の量のツイートがあれば経路を抽出可能かを評価する．

再現性が比較的良好であった池袋駅から大塚駅間の 9 領域を対象に，2〜4 日，7 日，1 カ月，3 カ月，6 カ月，9 カ月，12 カ月の期間のデータに対して本手法を適用して算出した距離誤差 RMSE と移動ツイート数の変動を**図 8** に示す．横軸は期間，左の縦軸は距離誤差 RMSE，右は領域から得られた移動ツイート数を表す．

全体的に，期間が長くなるにつれて移動ツイート数は増大し，揺らぎはあるが RMSE が減少することが確認できる．

また，7 日から 12 カ月までは RMSE は 18 m 未満と比較

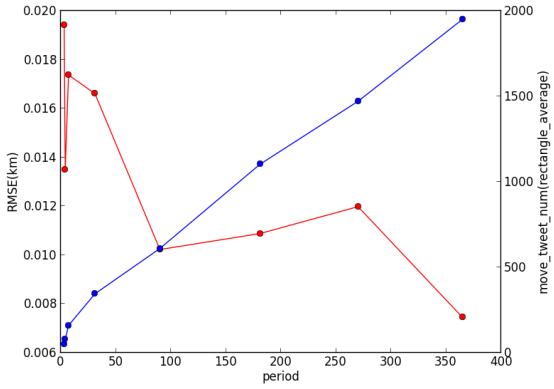


図 8　ある期間における RMSE と移動ツイート数の変動
Fig. 8　Variation of RMSE and number of tweets for each period.

的小さく，抽出された領域数は，1 カ月で領域の境界線上の路線が再現されておらず 8 領域であったが，他の期間はすべて抽出されていた．一番データ量が少ない 7 日間の場合は，移動ツイート数が 160，領域最小移動ツイート数が 5 であった．

ただし，4 日間の場合には，RMSE は同様に 13 m で小さかったが，再現された領域数は 6 であった．再現されていない領域は渋谷駅から大塚駅間の比較的曲率が大きい 3 つの領域であった．この理由は，曲率部では Hough 空間上の異なるグリッドに割り振られるからだと考えられる．ただし，スプライン補間により再現できた領域が 1 つ確認された．

3 日間の場合には，RMSE は 19 m と大きく，再現された領域は 2 つであり，2 日間の場合には再現されなかった．2 日間の場合の移動ツイート数は 30 で，そもそもツイートが存在しない領域は 2 つ確認できた．

4.9　動的に生成される経路抽出への応用

本論文では，提案手法を静的な鉄道路線データを用いて評価したが，すでに述べたとおり，提案手法に入力として与えるジオタグ付きツイートの選別方法を目的に合わせて変更することで，同じ目的を持つ多くのユーザが通ることで動的に生成されるような経路の発見に適用できる．たとえば，花見の時期には，多くのユーザは桜並木を見ながら，写真を撮影したり，そのときの状況をツイートしたりすることが考えられることから，実際に花見の時期に桜並木に沿って移動する花見客の移動経路が抽出できるかを検証した．

本文に「花見」，「はなみ」，「桜」，「サクラ」，「さくら」の文字列が含まれているツイートと，さらに同様に花見に関連する可能性が高い前後 1 個のツイートを抽出して，移動ツイート集合の代わりに提案手法の入力として与えた．ただし，ジオタグが GPS などの手段で測定したユーザの位置であることを保証するために，利用クライアントは

図 9　造幣局本局周辺の可視化結果
Fig. 9　Visualization result surrounding Japan mint head office.

「Twitter for Android」，「Twitter for iPhone」に限定した．また，本来は性能向上のために，パラメータを目的に合わせて調節することが望ましいが，提案手法の可能性を明確にするために，あえて 4.2 節と同じ値を使用した．

大阪市北区にある造幣局本局周辺の抽出経路の可視化結果を，**図 9** に示す．造幣局本局では，毎年春に「桜の通り抜け」と称して，敷地内の桜並木を観賞できるように，図 9 に青の点線で示す南門から北門までの約 560 m の経路を公開している．この敷地内の経路部分は，曲率が比較的大きいながらも比較的忠実に再現されていた．さらに，南門から天満橋駅の間にも，大川を斜めに横断するような形で経路が抽出されていた．桜の通り抜けは南門から北門への 1 方向であることから，天満橋駅で降りて造幣局の南門に向かう花見客が多いが，今回用いた矩形領域が大きすぎたことから，その天満橋のクランク状の経路が，緩やかな曲線として抽出されてしまったと考えられる．これ以外にも，中目黒駅周辺でも，中目黒駅から目黒川沿いに同様に経路が抽出されていた．

以上の結果から，多くのユーザが共同で利用する経路部分だけが比較的正しく抽出できることが分かる．ただし，路線抽出の場合と同様に，経路形状が複雑な部分の抽出性能を上げるためには，矩形領域の縮小や，アルゴリズムの改良が必要である．

上野公園や新宿御苑のような公園全体に桜がある場所では，うまく経路を抽出できなかった．これは桜や花見客が敷地内に広く分散しているからだと考えられる．このように，ユーザの移動ではなく滞在を分析したい場合は，DBSCAN や Mean Shift Clustering などの手法を用いて密度が高い領域として抽出する方が適切だと考えられる．

5. おわりに

本論文では，特に何らかの交通手段を利用しているときのジオタグ付きツイートから，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手法を提案した．実際に，JR 山手線周辺区域に本手法を適用し，路線ごとに Precision と Recall を評価し，ツイート数の減少にともなう RMSE を算出することで，本手法の有効性を示した．さらに，動的に生成される経路が抽出可能であることを示すため，桜並木に沿って移動する花見客の経路を分析した．

今後は，Precision が低下する原因として，複数路線が乗り入れる主要駅周辺での経路形状の複雑さに問題があるため，単なる連続性のある経路の集合ではなく，乗り換え駅をノード，その間の経路をエッジとするグラフ構造として抽出することを検討している．

謝辞　本研究は JSPS 科研費 26330345 の助成を受けた．

参考文献

[1] 佐伯圭介，遠藤雅樹，廣田雅春，倉田陽平，横山昌平，石川博：外国人 Twitter ユーザの観光訪問先の属性別分析，第 7 回データ工学と情報マネジメントに関するフォーラム（DEIM2015）(2015).

[2] Sakaki, T., Okazaki, M. and Matsuo, Y.: Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors, *Proc. 19th International Conference on World Wide Web* (*WWW '10*), pp.851–860 (2010).

[3] Hada, Y., Kodama, N., Suzuki, T. and Meguro, K.: Road Information Sharing Using Probe Vehicle Data in Disasters, *Proc. 14th World Conference on Earthquake Engineering* (*WCEE2008*) (2008).

[4] Hada, Y., Suzuki, T. and Noda, I.: Utilization of Probe Vehicle Information in Disasters in Japan, *Proc. 15th World Conference on Earthquake Engineering* (*WCEE2012*) (2012).

[5] 山田直治，礒田佳徳，南　正輝，森川博之：GPS 搭載携帯電話を用いた移動経路履歴に基づく訪問地・経由地予測システム，情報処理学会研究報告ユビキタスコンピューティングシステム（UBI），Vol.2010-UBI-27, No.4, pp.1–8 (2010).

[6] Wang, Y., Zheng, Y. and Xue, Y.: Travel Time Estimation of a Path Using Sparse Trajectories, *Proc. 20th ACM SIGKDD international conference on Knowledge discovery and data mining* (*KDD '14*), pp.25–34 (2014).

[7] Wei, L.-Y., Zheng, Y. and Peng, W.-C.: Constructing popular routes from uncertain trajectories, *Proc. 18th ACM SIGKDD international conference on Knowledge discovery and data mining* (*KDD '12*), pp.195–203 (2012).

[8] Krumm, J.: Trajectory Analysis for Driving, *Computing with Spatial Trajectories*, chapter 7, Zheng, Y. and Zhou, X. (Eds.), pp.213–241, Springer (2011).

[9] Reddy, S., Mun, M., Burke, J., Estrin, D., Hansen, M. and Srivastava, M.: Using Mobile Phones to Determine Transportation Modes, *ACM Trans. Sensor Networks* (*TOSN*), Vol.6, No.2, pp.13:1–13:27 (2010).

[10] Stenneth, L., Wolfson, O., Yu, P.S. and Xu, B.: Transportation Mode Detection using Mobile Phones and GIS Information, *Proc. 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pp.54–63 (2011).

[11] Zheng, Y., Chen, Y., Li, Q., Xie, X. and MA, W.-Y.: Understanding Transportation Modes Based on GPS Data for Web Applications, *ACM Trans. Web* (*TWEB*), Vol.4, No.1, pp.1:1–1:36 (2010).

[12] 遠藤結城，数原良彦，戸田浩之，小池義昌：移動手段判定のための表現学習を用いた GPS 軌跡からの特徴抽出，第 7 回 Web とデータベースに関するフォーラム（WebDB Forum 2014）(2014).

[13] 石野亜耶，小田原周平，難波英嗣，竹澤寿幸：Twitter からの被災時の行動経路の自動抽出および可視化，第 18 回年次大会，言語処理学会，pp.907–910 (2012).

[14] 大森雅己，廣田雅春，石川　博，横山昌平：ソーシャルメディア上から収集したジオタグに基づく地理的特徴の抽出と評価，情報処理学会論文誌データベース，Vol.8, No.1, pp.1–16 (2015).

[15] Duda, R.O. and Hart, P.E.: Use of the Hough Transformation To Detect Lines and Curves in Pictures, *Comm. ACM*, Vol.15, No.1, pp.11–15 (1972).

[16] 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).

[17] 国土交通省国土政策局：国土数値情報（鉄道時系列）製品仕様書 第 1.2 版 (2014).

[18] Misra, P. and Enge, P.: 改訂 第 2 版 精説 GPS，松香堂書店 (2010).

谷 直樹

2015 年和歌山大学システム工学部情報通信システム学科卒業．和歌山大学大学院システム工学研究科修士課程在学中．ソーシャルメディア分析の研究に従事．

風間 一洋　（正会員）

1988 年京都大学大学院工学研究科精密工学専攻修士課程修了．同年日本電信電話（株）入社．2005 年京都大学大学院情報学研究科システム科学専攻博士課程修了．博士（情報学）．2012 年和歌山大学システム工学部教授，現在に至る．Web 情報検索，Web マイニングの研究に従事．人工知能学会，日本ソフトウェア科学会，日本データベース学会，ACM 各会員．

榊 剛史

2004 年東京大学工学部電子情報工学科卒業．2006 年同大学院修士課程修了．電力会社通信部門での勤務を経て，2009 年東京大学大学院博士課程入学．2014 年博士課程修了．博士（工学）．東京大学での特任研究員を経て，2015 年より株式会社ホットリンク開発本部開発研究グループマネージャならびに東京大学客員研究員．専門は，自然言語処理，Web マイニング，社会ネットワーク分析．

吉田 光男　（正会員）

2009 年筑波大学第三学群情報学類卒業．2011 年同大学院システム情報工学研究科コンピュータサイエンス専攻博士前期課程修了，2014 年同博士後期課程修了．博士（工学）．同年より豊橋技術科学大学大学院工学研究科（情報・知能工学系）助教．ウェブ工学，自然言語処理，計算社会科学に関する研究に従事．言語処理学会，人工知能学会，日本データベース学会各会員．

斉藤 和巳　（正会員）

静岡県立大学経営情報学部教授．1985 年慶応義塾大学理工学部数理科学科数学専攻卒業，1998 年東京大学博士（工学）．複雑ネットワークの研究に従事．電子情報通信学会，人工知能学会，日本神経回路学会，日本応用数理学会，日本行動計量学会，日本データベース学会各会員．

（担当編集委員　土方 嘉徳）