

少量の移動履歴を考慮した大規模な集計データからの人流推定

Estimating People Flow from a Large Amount of Aggregated Data with a Few Tracking Data

河井 悠佑[♡] 田中 佑典[◇]
戸田 浩之[♣] 石川 佳治[♣]

Yusuke KAWAI Yusuke TANAKA
Hiroyuki TODA Yoshiharu ISHIKAWA

人の位置情報は、プライバシー保護が必要なデータである。そのため、追跡情報を含んだ個人の移動履歴の取得は少数のユーザに限られることが多い。一方で、エリア人数のような集計化されたデータは大規模に取得しやすい。本論文では、実用的なデータ計測環境を想定し、少量の移動履歴と大規模な集計データとを考慮した高精度な人流推定手法を提案する。提案手法では、移動履歴中の追跡可能なユーザと集計データに含まれるユーザが、共通の遷移確率に従って移動を行うという仮説に基づく確率モデルを学習する。実データに基づく実験により、提案手法が観測点間の人流を高精度に推定できることを示した。また、追跡可能なユーザ数と推定精度の関係を検証した。

Privacy concerns for location information are recently increasing. This makes it difficult to gather a large amount of tracking data, i.e., histories of location and time for each individual, and thus we often obtain the tracking data for the limited number of people. In contrast, we can easily obtain a large amount of aggregated data, such as population in a certain area; they do not contain tracking information of individuals. In this paper, assuming a practical data acquisition environment, we propose a people flow estimation method by combining two types of data, i.e., a few tracking data and a large amount of aggregated data. The proposed method assumes that people move from location to location in accordance with the transition probabilities that is shared across both datasets. By maximizing the likelihood with flow conservation constraints, our model estimate the people flows between locations via utilizing two types of data. The experimental results using real-world

datasets demonstrate the performance improvement of our model. We also examine the relationship between the number of trackable individuals and estimation accuracy.

1 はじめに

今日では、モバイル端末の普及やセンサ技術の発展により、人の位置情報の取得が容易になってきた。そのため、イベント会場やショッピングモールといった大量の来場者（以降、ユーザと呼ぶ）が滞在する場所において、広告配信 [7] やナビゲーション [6, 19] といったサービスへの応用が期待される。例えば、イベント会場において、ユーザが「どこからどこへ移動しやすいか」という移動の傾向が分かれば、人気のルートを提示することができる。また、ショッピングモールでは、ユーザが次に訪問しそうな店舗を分析し、効果的な広告配信が可能である。

人の位置情報は、個人を追跡可能な移動履歴と、エリア内の人口のように集計化された集計データの二つに大別される。以下では、これらが持つメリットとデメリットについて述べる。個人の移動履歴とは、個人単位で時刻と所在地が逐次記録されたものであり、「誰が」「いつ」「どこから」「どこへ」移動したか把握できる情報となっている。取得方法としては、屋外では GPS が活用されていることが多く、屋内では Wi-Fi 基地局や BLE ビーコンとユーザの端末との通信を利用することが検討されている [5, 16]。しかしながら、移動履歴データをまとめて分析するには、個々のユーザの端末に、データを収集してサーバに送信するアプリ等の仕掛けを組み込むことが必要となり、大量のユーザの情報を取得することはコストの問題で難しい。それに加えて、個人の移動履歴は、個人に関するプライバシー性の高い情報であるため、利活用に対するユーザの不安が大きい。総務省情報通信政策研究所¹が行った「位置情報の利用に関する意識調査」[1]では、企業による位置情報サービスの利用に関して、50%以上が条件付きでの許容、20%以上が許容できない、としている。したがって、個人の移動履歴の取得について、多くのユーザから許諾を得られにくいのが現状である。以上の点から、イベント会場における個人の移動履歴データの取得量は、全体の数%のユーザに留まることが多い。その場合、イベント会場全体の移動の流れを高精度にとらえることは困難である。

一方で、集計データとは、ある時刻と場所に滞在する人数を、集計化して計測したものであり、「いつ」「どこに」「何人」いたかを把握できる情報である。本論文では、一定時間において観測領域内から出た人の総数（以降、流出人数と呼ぶ）と入ってきた人の総数（以降、流入人数と呼ぶ）が観測されていると想定する。個人単位の情報は保持していないため、ユーザのプライバシーを保護したまま分析することができる。さらに、個人の追跡が不要であるため、カメラやレーザ等の設置が比較的容易なセンサによる計測が可能であり、個々のユーザにアプリのインストール等の手間を取らせる必要はない。また、個人の移動履歴として取得したデータも、プライバシー保護のために集計化して扱う場合が想定される [15]。したがって、集計データは取得コストやプライバシーの問題を緩和できるため、個人の移動履歴と比べて大規模に取得可能である。しかし、追跡情報が失われているため、「どこから」「どこへ」移動したかは把握できない。

以上で述べたデータ計測の問題を踏まえると、個人の移動履歴は移動情報は保持しているものの少量しか取得できず、集計データは大規模に取得できるものの移動に関する情報が失われている。したがって、本論文で少量の個人の移動履歴と大規模な集計データが取得可能な状況を想定することは、現実的な背景を踏まえた自然かつ実用的な問題設定である。例えばショッピングモールでは、店ごとにセンサを設置して多数のユーザの人数を集計をしつ

[♡] 学生会員 名古屋大学情報学研究科
kawai@db.is.i.nagoya-u.ac.jp

[◇] 非会員 日本電信電話株式会社 サービスエボリューション研究所
tanaka.y@lab.ntt.co.jp

[♣] 正会員 日本電信電話株式会社 サービスエボリューション研究所
toda.hiroyuki@lab.ntt.co.jp

[♣] 正会員 名古屋大学情報学研究科
ishikawa@i.nagoya-u.ac.jp

この研究は第 1 著者が NTT サービスエボリューション研究所でのインターンシップ期間中に実習生として行ったものである。

¹ <http://www.soumu.go.jp/iicp/>

つ、協力を頂いたユーザからは店から店への移動履歴が取得することが考えられる。

集計データのみから移動人数を推定する手法がいくつか提案されている [2, 8, 9, 11, 22]。例えば、Collective Flow Diffusion Model (CFDM) [11] では、時間定常な移動確率に従って各ユーザが移動すると仮定し、各時間ステップにおいて移動人数の総量が保存される、という制約を加味した最適化問題を解くことによって、流出人数と流入人数のみを入力データとして人流の推定を行う。しかしながら、入力データには「どこから」「どこへ」移動したかという情報が一切含まれていない。そのため、実際の人流とはかけ離れた推定結果を出力してしまう場合がある。例えば、混雑回避による移動規制や遠く離れた場所といった、現実的に困難な移動状況であろうとも、移動元の場所における流出人数と移動先の場所における流入人数がともに大きければ、その間の移動人数は大きく推定されてしまう。

本研究では、データ計測の実用的な状況を踏まえて、少量の移動履歴と大規模な集計データから高精度に人流推定を行う手法を提案する。集計データと個人の移動履歴という二つの異なる形式から、イベント会場における移動のモデル化及び移動人数の推定を行う。提案手法では、移動履歴中の追跡可能なユーザと集計データに含まれるユーザが、共通の移動確率に従って移動を行うという仮説に基づく確率モデルを学習する。

実験では実際のイベント会場におけるデータを用いて、従来手法及び本手法の推定精度を比較することにより本手法の有用性を示す。また、個人の移動履歴の量を変えて実験を行い、データ量と推定精度との関係を検証する。その結果、少量の個人の移動履歴と集計データを組み合わせたときに、精度の大きな向上が見られ、本手法の有効性を示すことができた。

本論文の貢献を以下に示す。

- 集計データと個人の移動履歴を同時に考慮した移動推定モデルの提案
- 実験による提案手法の有用性の検証
- 個人の移動履歴の量と推定精度との関係の検証

本論文の構成は以下の通りである。まず、第 2 章で関連研究について述べる。次に、第 3 章で基本的な概念や問題設定について述べる。第 4 章で従来手法である CFDM を紹介し、それを踏まえて第 5 章で CFDM をベースとした提案手法について述べる。第 6 章で提案手法の有用性を確認する実験を行い、最後に第 7 章で本論文のまとめと今後の課題について述べる。

2 関連研究

本章では、位置情報から生成されたデータを用いた人流の推定手法に関する研究について述べる。位置情報によって生成されるデータの代表的なものとして、個人の移動履歴と集計データがあり、それぞれのデータに基づいた手法について述べる。

個人の移動履歴に基づいた移動傾向の分析手法としてマルコフモデルがある。マルコフモデルは、現在の位置から次の位置への移動が決定される、というマルコフ性を仮定した確率モデルを構築したものである。個人の移動履歴の「どこからどこへ移動した」という情報を用いて、観測点間の移動確率を求めることができる。この移動確率から、現在の位置が与えられたユーザの次の行動の予測が可能になる。このような研究は、災害管理 [20]、交通 [10, 19]、観光 [12]、SNS [4] などの様々な分野で行われている。しかしながら、今回対象とするイベント会場などにおいては、必ずしも十分にデータを取得することができず、ユーザ全体の移動傾向をとらえることは難しい。また、個人の移動履歴と比べて大規模に取得可能である集計データがあるが、「どこからどこへ移動した」という情報が失われているため、この手法において活用できない。提案手法は、個人の移動履歴も集計データも活用した推定手法である。それに加えて、個人の移動履歴が少量の時でも、精度良く移動確率の推定が行える。

表 1: 本研究で扱う記号

記号	説明
V	観測領域の集合
i	観測領域, $i \in V$
E_i	観測領域 i の近傍の観測領域の集合
j	観測領域 i の近傍の観測領域, $j \in E_i$
T	観測時間ステップの上限
t	観測時間ステップ, $t \in \{1, \dots, T\}$
N_{ii}^{OUT}	時刻 t に i から流出する集計ユーザ数
N_{ii}^{IN}	時刻 t に i へ流入する集計ユーザ数
R_{ii}^{OUT}	時刻 t に i から流出する個人ユーザ数
R_{ii}^{IN}	時刻 t に i へ流入する個人ユーザ数
L_{tij}	時刻 t に i から j へ移動する個人ユーザ数
θ_{ij}	i から j へ移動する確率 ただし, $\theta_{ij} \geq 0, \sum_{j \in E_i} \theta_{ij} = 1$ を満たす

集計データから移動人数を推定する手法に関する研究が盛んである [2, 8, 9, 11, 22]。Collective Flow Diffusion Model (CFDM) [11] はその一つであり、集計データから観測点間の移動人数を表す潜在変数と観測点間の移動確率モデルを定義したものである。移動人数が保存する制約を加味した最適化問題を解くことによって、パラメータを推定する。元の論文では、交通ネットワークにおける交通流の推定を対象としている。交差点から出て行った車両数と交差点に入ってきた車両数から、交差点における方向転換の確率を推定する。この手法の詳細については第 4 章で述べる。また、このモデルの改良として、Variational Bayesian Collective Graphical Mixture Model (VCGMM) [9] が提案されている。これは、通勤などの時間帯による移動の様子の違いを考慮して、それぞれの時間帯に応じた移動確率を推定できる。CFDM や VCGMM は、エリア内の人数といった集計データのみを入力として、移動確率の推定を行っている。提案手法は、この集計データに加えて、ユーザの移動履歴も合わせて確率モデルを構築した手法となっている。

集計データを分析する一般的なフレームワークとして、Collective Graphical Model (CGM) [18] が提案されている。これまでの研究において、CGM におけるパラメータの推定手法として、マルコフ連鎖モンテカルロ法 [18]、最大事後確率推定 [17]、ガウス近似 [13]、確率伝播法 [21]、DC (difference of convex functions) 計画問題 [14] といった様々なものが提案されている。CFDM は CGM の特別な場合であり、オブジェクトの流れを推定するのに適したモデルとなっている。したがって、CFDM は CGM と同様の推定手法を適用することができる。本研究では、[17] で提案されている効率的な手法を採用する。

3 問題設定

本章では、本論文で扱う人の移動に関して、基本的な概念や問題設定について述べる。本論文で扱う記号を表 1 に示す。

本研究が想定する設定について述べる。状況は、イベント会場などの有限空間におけるユーザの移動を想定する。ブースなどのスポットとなる場所を観測領域として、集計データ及び個人の移動履歴を計測可能な機器を設置する。集計データとして取得されているユーザを集計ユーザ、個人の移動履歴として取得されているユーザを個人ユーザと呼ぶ。集計ユーザと個人ユーザは分離されているものとし、集計ユーザと個人ユーザそれぞれの変数を定義する。

3.1 データ

まず、観測する領域と時間について述べる。観測領域の集合を V とし、その要素である観測領域を $i \in V$ と表す。観測領域 i の近傍の観測領域の集合を E_i とし、その要素である近傍の観測領域を $j \in E_i$ と表す。全観測時間のある単位時間で分割し、その時間ステップの一つを観測時間ステップ t とする。最初の時間ステップを 1, 上限を T とし、 $t \in \{1, \dots, T\}$ と表す。

次に、集計ユーザに関して、計測機器により観測可能な変数（以下、観測変数とする）について述べる。 i における t ごとの流出人数 N_i^{OUT} と流入人数 N_i^{IN} は、それぞれ t において i から出ていった集計ユーザの総数と、入ってきた集計ユーザの総数を表す。それらをまとめて $N^{OUT} = \{N_i^{OUT} | t = 1, \dots, T, i \in V\}$, $N^{IN} = \{N_i^{IN} | t = 1, \dots, T, i \in V\}$ として、 t, i の二次元配列で表現する。

個人ユーザに関する観測変数について述べる。流出人数と流入人数は集計ユーザと同様の形式であり、それぞれ R_i^{OUT}, R_i^{IN} と表す。これらをまとめて二次元配列 $R^{OUT} = \{R_i^{OUT} | t = 1, \dots, T, i \in V\}$, $R^{IN} = \{R_i^{IN} | t = 1, \dots, T, i \in V\}$ として表現する。 t ごとの i から j への個人ユーザの移動人数を L_{ij} と表す。これをまとめて $L = \{L_{ij} | t = 1, \dots, T, i \in V, j \in E_i\}$ として、 t, i, j を次元とする三次元配列で表現する。

3.2 問題

観測変数 $N^{OUT}, N^{IN}, R^{OUT}, R^{IN}, L$ を踏まえて、 i から j へ移動する確率 θ_{ij} を推定する。この移動確率は、観測領域のみに依存し、全ユーザが共通の移動確率に従って移動するという仮説に基づいたものである。そのため、集計ユーザと個人ユーザに共通のパラメータとなっている。これをまとめて $\theta = \{\theta_{ij} | i \in V, j \in E_i\}$ として、 i, j を次元とする二次元配列で表現する。

本研究では、集計ユーザの流出人数 N^{OUT} と流入人数 N^{IN} 、個人ユーザの流出人数 R^{OUT} と流入人数 R^{IN} 及び移動人数 L を入力とする。入力から移動確率 θ を推定して出力する。

4 Collective Flow Diffusion Model

本章では、集計データのみを用いて移動推定を行う従来手法 Collective Flow Diffusion Model (CFDM) [11] を紹介する。元の論文では、道路ネットワークにおける車両の移動を対象としていたが、本論文では人の移動に適用する。集計ユーザの流出人数 N^{OUT} と流入人数 N^{IN} を入力とする。この二つの観測変数から、ユーザの移動確率 θ を推定して出力する。CFDM では、移動に関する確率モデルと制約を定義する。この確率モデルと制約に関する最適化を行うことにより、パラメータ θ の最適な値を推定する。

4.1 確率モデル

人の移動に関する確率モデルを尤度関数として定義する。この定義に当たって、集計ユーザの移動人数 M_{ij} を潜在変数として導入する。 t において i から j へ移動した集計ユーザの総数であり、形式は L_{ij} と同じである。これをまとめて $M = \{M_{ij} | t = 1, \dots, T, i \in V, j \in E_i\}$ として、 t, i, j を次元とする三次元配列で表現する。 N^{OUT} が得られたとき、移動確率 $\theta_i = \{\theta_{ij} | j \in E_i\}$ の多項分布に従って $M_i = \{M_{ij} | j \in E_i\}$ が得られたと仮定する。このとき M_i の尤度関数は以下となる。

$$P(M_i | N_i^{OUT}, \theta_i) = \frac{N_i^{OUT}!}{\prod_{j \in E_i} M_{ij}!} \prod_{j \in E_i} \theta_{ij}^{M_{ij}}. \quad (1)$$

次に、 N^{OUT}, N^{IN} と M に関して制約を設ける。「 i を出発したユーザは必ず $j \in E_i$ へ移動する」という移動人数の保存則を仮定する。つまり、移動は ij 間でのみ行われており、 i から $j' \notin E_i$ への移動や、観測領域外からの移動は想定しない。この仮定により、 i から移動した人数が i における流出人数に対応し、 i へ移動した人数が i における流入人数に対応する。

$$N_i^{OUT} = \sum_{j \in E_i} M_{ij} \quad (2)$$

$$N_i^{IN} = \sum_{j \in E_i} M_{tji}. \quad (3)$$

したがって、各時刻における M を行列としたとき、行の和が N^{OUT} 、列の和が N^{IN} となる。

4.2 推定

式 (1)(2)(3) からパラメータ θ の推定を行う。尤度関数と制約から目的関数を定めて、潜在変数 M とパラメータ θ が最適な値をとるように推定する。

4.2.1 目的関数

まず、式 (1) について対数をとることにより、対数尤度関数

$$\begin{aligned} \log P(M_i | N_i^{OUT}, \theta_i) \\ = \log N_i^{OUT}! - \sum_{j \in E_i} \log M_{ij}! + \sum_{j \in E_i} M_{ij} \log \theta_{ij} \\ \approx \sum_{j \in E_i} [(1 + \log \theta_{ij}) M_{ij} - M_{ij} \log M_{ij} - M_{ij}] \end{aligned} \quad (4)$$

が得られる。導入した潜在変数 M と求めたいパラメータ θ に関する項 $\log N_i^{OUT}!$ を削除して、スターリングの近似 $\log M_{ij}! \approx M_{ij} \log M_{ij} - M_{ij}$ を行うことによって、対数尤度関数を近似している [17]。そして、制約 (2),(3) を損失項として加えることによって、 M, θ に関する目的関数は以下となる。

$$\begin{aligned} \mathcal{L}(M, \theta) \\ = \sum_{t=1}^T \sum_{i \in V} \left(\sum_{j \in E_i} [(1 + \log \theta_{ij}) M_{ij} - M_{ij} \log M_{ij} - M_{ij}] \right. \\ \left. - \frac{\lambda}{2} |N_i^{OUT} - \sum_{j \in E_i} M_{ij}|^2 - \frac{\lambda}{2} |N_i^{IN} - \sum_{j \in E_i} M_{tji}|^2 \right). \end{aligned} \quad (5)$$

実世界の移動モデルでは移動人数の保存則を厳密に満たすとは限らないため、制約の強さを示すハイパーパラメータ λ を定めてソフトな制約とする。 λ は入力データに基づいて最適な値を決定する。この決定方法については第 4.2.3 節で述べる。

4.2.2 目的関数の最適化

目的関数 (5) は尤度関数が大きくなるほど、また制約を満たすほど値は大きくなる。したがって、(5) が最大になるような M, θ を求めたい。ここでは、二つの変数に関して、片方の変数を固定しつつ最大化することにより更新を行う。値が収束するまで更新を交互に繰り返すことにより、それぞれの最適な値を推定する。

まず、 M の更新について述べる。 M は移動人数であるため、全ての要素が整数である必要があり、目的関数 (5) に関する整数計画問題を解く必要がある。しかし、整数計画問題を解くためには非常に時間がかかってしまうため、 M を実数とした凸最適化問題とすることによって、計算時間を削減する。本研究では、準ニュートン法の一つである L-BFGS-B 法 [3] を用いて、式 (5) の値を最大にする。以下の最適化問題を解くことによって、 M の推定値が得られる。

$$\begin{aligned} \underset{M}{\text{maximize}} \quad & \mathcal{L}(M, \hat{\theta}) \\ \text{subject to} \quad & M_{ij} \geq 0, t = 0, \dots, T-1, i \in V, j \in E_i. \end{aligned} \quad (6)$$

ただし、 $\hat{\theta}$ は θ の現在の推定値を表す。

次に、 θ の更新について述べる。まず、式 (5) から θ に関する項だけを取り出すことによって、以下の式が得られる。

$$\begin{aligned}\mathcal{L}(\hat{\mathbf{M}}, \theta) &\propto \sum_{t=1}^T \sum_{i \in V} \sum_{j \in E_i} M_{tij} \log \theta_{ij} \\ &\propto M_{\cdot ij} \sum_{i \in V} \sum_{j \in E_i} \log \theta_{ij}.\end{aligned}\quad (7)$$

ただし、 $\hat{\mathbf{M}}$ は $\mathbf{M}$ の現在の推定値を表す。また、 $M_{\cdot ij} = \sum_{t=1}^T M_{tij}$ とした。式 (7) と制約条件 $\sum_{j \in E_i} \theta_{ij} = 1$ から、ラグランジュの未定乗数法を用いる。ラグランジュ乗数を η として θ に関する目的関数は以下となる。

$$\mathcal{J}(\theta) = M_{\cdot ij} \sum_{i \in V} \sum_{j \in E_i} \log \theta_{ij} + \eta \sum_{j \in E_i} (\theta_{ij} - 1). \quad (8)$$

目的関数 (8) が最大になるとき、 θ_{ij} に関する偏微分が 0 となるため、以下の式が得られる。

$$\begin{aligned}\frac{\partial \mathcal{J}(\theta)}{\partial \theta_{ij}} &= \frac{M_{\cdot ij}}{\theta_{ij}} + \eta = 0 \\ \Leftrightarrow -\frac{1}{\eta} M_{\cdot ij} &= \theta_{ij}.\end{aligned}\quad (9)$$

両辺に $\sum_{j \in E_i}$ をとると、制約条件 $\sum_{j \in E_i} \theta_{ij} = 1$ より η は

$$\eta = -\sum_{j \in E_i} M_{\cdot ij} \quad (10)$$

となる。したがって、 θ_{ij} は以下のように求められる。

$$\theta_{ij} = \frac{M_{\cdot ij}}{\sum_{j \in E_i} M_{\cdot ij}}. \quad (11)$$

4.2.3 λ の決定

ハイパーパラメータ λ は、実際のデータにおける観測変数 N^{IN}, N^{OUT} に基づいて最適な値を決定する。まず、 λ の候補になる値をいくつか用意する。それぞれの λ について、第 4.2.2 節をもとに移動確率 $\hat{\theta}_{ij}$ を推定する。暫定の移動確率 $\hat{\theta}_{ij}$ から流入人数を

$$\hat{N}_i^{IN} = \sum_{j \in E_i} \hat{\theta}_{ji} N_{ij}^{OUT} \quad (12)$$

と推定する。 λ ごとの推定流入人数 $\hat{N}_i^{IN}$ と実際の流入人数 N_i^{IN} との誤差を計算し、最小のときの λ をこのモデルにおける最適な λ とする。

5 提案手法

本章では、第 4 章を踏まえて、提案手法について述べる。集計データに加えて、個人の移動履歴を考慮してパラメータの推定を行う。提案手法では、 N^{OUT} と N^{IN} に加えて、個人ユーザの移動人数 $\mathbf{L}$ を入力とする。推定対象は従来手法と同じ集計ユーザであり、ユーザの移動確率 θ を出力する。

5.1 確率モデル

提案手法では、集計ユーザの移動人数 $\mathbf{M}$ と個人ユーザの移動人数 $\mathbf{L}$ とが、共通の移動確率 θ に従って生成されると仮定する。このとき、 $\mathbf{M}$ と $\mathbf{L}$ の同時確率は

$$\begin{aligned}P(\mathbf{M}_{ii}, \mathbf{L}_{ii} | N_{ii}^{OUT}, R_{ii}^{OUT}, \theta_i) \\ = \frac{N_{ii}^{OUT}!}{\prod_{j \in E_i} M_{tij}!} \prod_{j \in E_i} \theta_{ij}^{M_{tij}} \frac{R_{ii}^{OUT}!}{\prod_{j \in E_i} L_{tij}!} \prod_{j \in E_i} \theta_{ij}^{L_{tij}}\end{aligned}\quad (13)$$

と表される。

Algorithm 1 推定アルゴリズム

Input: $N^{OUT}, N^{IN}, \mathbf{L}$

Output: $\mathbf{M}, \theta$

initialize ϵ

while $\epsilon \leq \epsilon$ **do**

 update M_{tij} by maximizing (5)

 update θ_{ij} by (16) for $i \in V, j \in E_i$

$\epsilon \leftarrow \sum_{t=1}^T \sum_{i \in V} \sum_{j \in E_i} |M_{tij} - M_{tij}^{old}|$

$M_{tij}^{old} \leftarrow M_{tij}$

end while

5.2 推定

式 (13) を尤度としたときの $\mathbf{M}, \theta$ に関する目的関数は以下となる。

$$\begin{aligned}\mathcal{L}(\mathbf{M}, \theta) \\ = \sum_{t=1}^T \sum_{i \in V} \left(\sum_{j \in E_i} [(1 + \log \theta_{ij}) M_{tij} - M_{tij} \log M_{tij} - M_{tij}] \right. \\ \left. - \frac{\lambda}{2} |N_{ii}^{OUT} - \sum_{j \in E_i} M_{tij}|^2 - \frac{\lambda}{2} |N_{ii}^{IN} - \sum_{j \in E_i} M_{tij}|^2 \right) \\ + \sum_{t=1}^T \sum_{i \in V} \sum_{j \in E_i} (\log \theta_{ij}) L_{tij}.\end{aligned}\quad (14)$$

式 (5) に、個人ユーザにおける θ に関する項 $(\log \theta_{ij}) L_{tij}$ を加えた形となる。第 4.2.2 節と同様に、式 (14) が最大になるように $\mathbf{M}$ と θ の更新を繰り返す。式 (14) は、 $\mathbf{M}$ に関する項が式 (5) と同じである。そのため、 $\mathbf{M}$ の更新は式 (6) と同じである。式 (14) から、 θ についての項のみを取り出すと、 θ の目的関数は

$$\begin{aligned}\mathcal{G}(\theta) &= M_{\cdot ij} \sum_{i \in V} \sum_{j \in E_i} \log \theta_{ij} + L_{\cdot ij} \sum_{i \in V} \sum_{j \in E_i} \log \theta_{ij} \\ &= (M_{\cdot ij} + L_{\cdot ij}) \sum_{i \in V} \sum_{j \in E_i} \log \theta_{ij}\end{aligned}\quad (15)$$

となる。ここで、 $L_{\cdot ij} = \sum_{t=1}^T L_{tij}$ とした。ラグランジュの未定乗数法により、 θ_{ij} は以下となる。

$$\theta_{ij} = \frac{M_{\cdot ij} + L_{\cdot ij}}{\sum_{j \in E_i} (M_{\cdot ij} + L_{\cdot ij})}. \quad (16)$$

したがって、式 (11) を式 (16) に変更して従来手法と同様に $\mathbf{M}, \theta$ を推定すればよい。これを踏まえて、 $\mathbf{M}, \theta$ を求めるアルゴリズム 1 を示す。 $\mathbf{M}$ が収束するまで $\mathbf{M}$ と θ の最大化を交互に繰り返す。収束条件として ϵ を定義し、更新した推定値 $\mathbf{M}$ と一つ前のステップで求めた $\mathbf{M}^{old}$ との差 ϵ が ϵ 以下となったところで収束とする。このときの $\mathbf{M}$ と θ を最終的な推定値とする。

6 実験

提案手法の有効性を検証するために、実際のイベント会場におけるデータを用いて行った実験について述べる。実験の目的は以下の二つである。

- 提案手法と従来手法の比較
- 個人の移動履歴の量と推定精度の関係の検証

比較する従来手法は、第 4 章で述べた集計データのみから推定する CFDM と、個人の移動履歴のみから推定するマルコフモデルである。マルコフモデルは、現在の観測地点のみから移動先を決定する単純マルコフモデルとする。マルコフモデルでは、個人

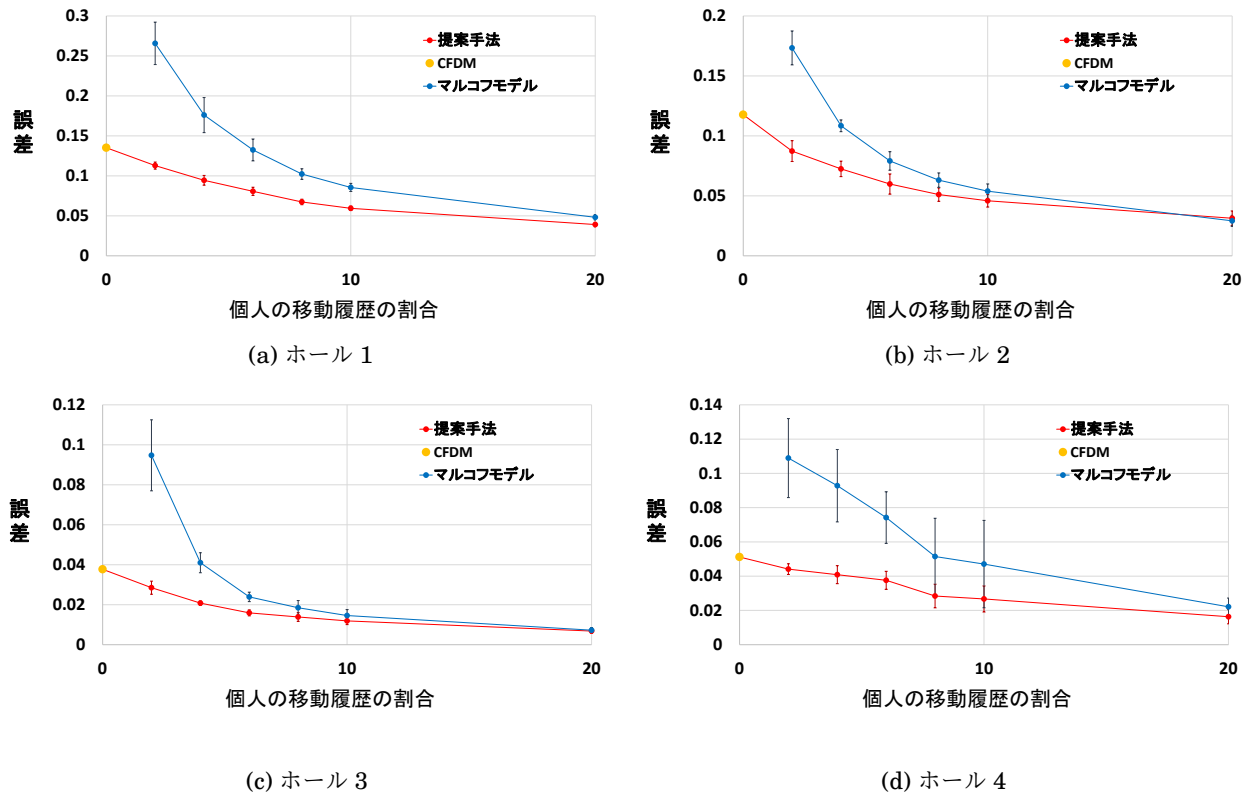


図 1: 各手法の誤差。横軸は移動履歴の割合、縦軸は真の移動確率と推定した移動確率との Jensen-Shannon ダイバージェンスである。小さいほど推定精度が良いことを示す。提案手法とマルコフモデルでは、5 回の実験における平均値をプロットし、誤差棒は標準誤差を示す。CFDM は、提案手法における移動履歴の割合が 0% の点（オレンジ）に対応する。提案手法もマルコフモデルも個人の移動履歴の割合が大きいくほど、誤差が小さくなっていき、推定精度が良くなることが分かる。

ユーザの移動人数 L を用いて、遷移確率 θ を以下の式で求める

$$\theta_{ij} = \frac{L_{ij}}{\sum_{j \in E_i} L_{ij}}. \quad (17)$$

なお、マルコフモデルで推定するパラメータは遷移確率 θ のみである。そのため、各手法の推定精度の比較は θ を用いて行う。

また、集計データに対する、個人の移動履歴に含まれるユーザ数の割合（以下、個人ユーザ数の割合）を変えることによって、個人の移動履歴の量と推定精度の関係を検証する。割合を変えたときの、提案手法及びマルコフモデルの推定精度の変化の様子を検証する。本研究は実用的なデータ計測の状況を想定しているため、集計ユーザ数に対する個人ユーザ数の割合が小さいときの推定精度に着目する。

本章では、まず実験の準備として、実験で用いたデータセットについて説明する。次に、実験の設定について述べる。最後に定量評価と定性評価の結果をそれぞれ示し、提案手法の有効性について検証する。

6.1 データセット

本研究で使ったデータセットは、ニコニコ超会議 2016² という実際のイベントにおいて取得したものである。イベント会場における各ブースにビーコンを設置し、来場者の協力のもと、スマートフォンアプリを用いて午前 10 時から午後 6 時まで移動履歴を取得した。取得したデータのユーザ数は合計 3727 人で、各ユーザごとのブースの入場時刻と退場時刻が記録されている。イベン

ト会場は大きく分けてホール 1、ホール 2、ホール 3、ホール 4 の 4 つがあり、それぞれ会場ごとにデータセットも分割した。これを集計ユーザと個人ユーザに分割して集計し、 N^{OUT}, N^{IN}, L を作成した。各ホールのブース数は、ホール 1 が 38、ホール 2 が 27、ホール 3 が 10、ホール 4 が 9 である。観測時刻 t の単位は 5 分とし、このとき $T = 96$ である

6.2 実験設定

実験では、 N^{OUT}, N^{IN}, L を用いて、各手法における集計ユーザの移動人数 M と移動確率 θ を推定する。なお、マルコフモデルでは L を用いて移動確率 θ のみを推定する。今回の比較対象は θ であり、推定した移動と真の移動を区別するために、推定した移動確率を θ^{est} 、真の移動確率を θ^{true} とする。今回の推定は、会場ごとの移動の様子のみを対象として、会場をまたいだ移動は考慮しない。

実験対象は、集計データのみから推定する CFDM と、個人の移動履歴のみから推定するマルコフモデル、集計データと個人の移動履歴から推定する提案手法である。準備として、個人ユーザ数の割合を変えたときの、 N^{OUT}, N^{IN}, L を作成する。全ユーザ数に対する個人ユーザ数の割合を $a\%$ とし、残りを集計ユーザとする。例えば、ホール 1 における全ユーザが 1000 人、 $a = 10$ のとき、集計ユーザは 900 人、個人ユーザは 100 人となる。 $a = 0, 2, 4, 6, 8, 10, 20$ における N^{OUT}, N^{IN}, L を作成し、それぞれマルコフモデルと提案手法を適用して θ^{est} を推定した。なお、 $a = 0$ は集計データのみであるため、提案手法が CFDM に相当し、マルコフモデルは推定しない。真の移動確率 θ^{true} は、観測ユーザが全て追跡可能な個人

²<http://chokaigi.jp/2016/>

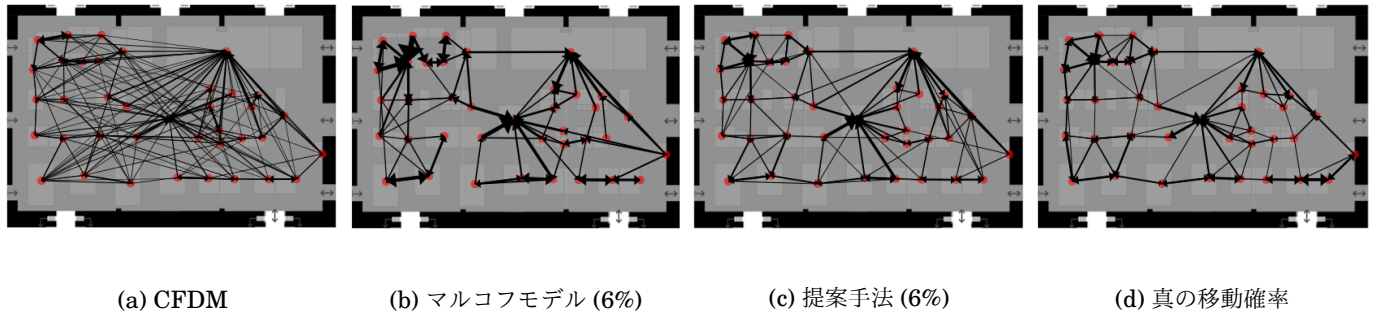


図 2: ホール 1 の移動の様子。赤の点は各ブースを示し、矢印はブース間の移動を示す。矢印の始点は移動元のブース、終点が移動先のブースである。移動確率の値が大きいほど、対応するブース間の矢印は太くなる。煩雑な図示を防ぐため、確率の低い移動は除外してある。双方向に頻繁に移動しているブース間は、双方向の矢印によって図示し、太さは二つの値の平均としている。

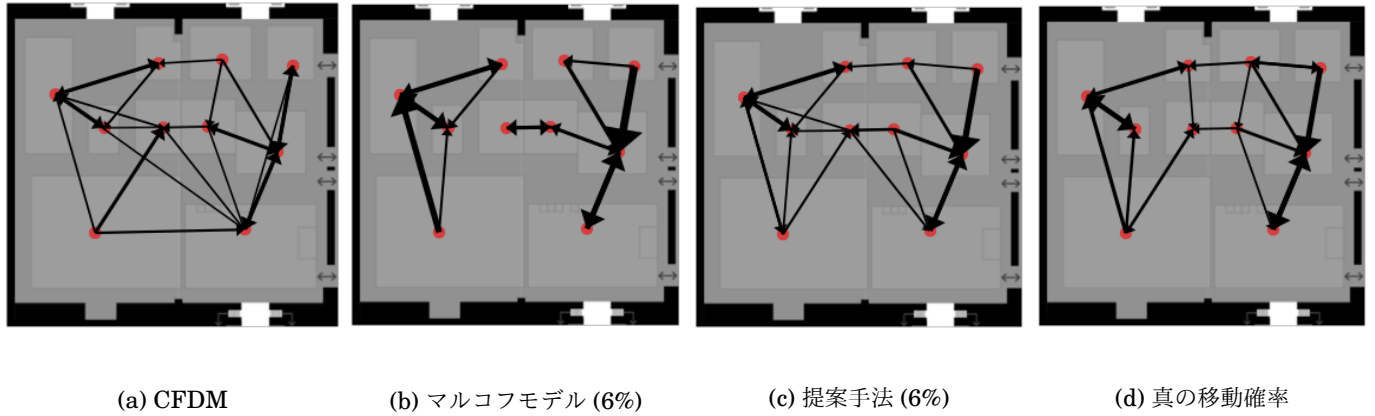


図 3: ホール 3 の移動の様子。ホール 1 の移動の様子（図 2）と同様に、赤の点は各ブースを、矢印はブース間の移動を示す。移動確率の値が大きいほど、対応するブース間の矢印は太くなる。

ユーザとしたときの移動確率とする。つまり、 $a = 100$ としたときの式 (17) を計算したものである。以降の結果の評価において、各手法における θ^{est} と θ^{true} を比較する。また、ユーザの偏りを考慮して、一つの割合 a に対して 5 分割で交差検証を行った。

6.3 定量評価

推定した移動確率 θ^{est} と真の移動確率 θ^{true} との誤差を計算して定量評価を行った。誤差の尺度には、Jensen-Shannon ダイバージェンスを用いた。これは、確率分布の差異の代表的な尺度である Kullback-Leibler ダイバージェンスを対称にしたものである。各 i の Jensen-Shannon ダイバージェンスは、以下の式で求められる

$$\begin{aligned}
 JS(\theta_i^{est} \parallel \theta_i^{true}) &= \frac{1}{2} KL(\theta_i^{est} \parallel \theta'_i) + \frac{1}{2} KL(\theta_i^{true} \parallel \theta'_i) \\
 &= \frac{1}{2} \sum_{j \in E_i} \theta_{ij}^{est} \log \frac{\theta_{ij}^{est}}{\theta'_{ij}} + \frac{1}{2} \sum_{j \in E_i} \theta_{ij}^{true} \log \frac{\theta_{ij}^{true}}{\theta'_{ij}} \quad (18) \\
 \text{ただし、} \theta'_i &= \frac{1}{2} (\theta_i^{est} + \theta_i^{true}).
 \end{aligned}$$

最終的な誤差の値は、各 i の Jensen-Shannon ダイバージェンスの平均とした。 θ^{est} が θ^{true} に近づくほど、Jensen-Shannon ダイバージェンスの値は小さくなるため、推定精度が良いといえる。実験設定で示した各手法について、それぞれ θ^{est} を求めて、 θ^{true} との Jensen-Shannon ダイバージェンスを計算した。

結果は図 1 のようになった。横軸は個人ユーザ数の割合 a 、縦軸を誤差の尺度である Jensen-Shannon ダイバージェンス（式 18）の値としている。縦軸の値が小さいほど推定精度は良いといえる。提案手法とマルコフモデルについて、5 回の実験における平均値をプロットし、標準誤差を誤差棒としている。提案手法における個人ユーザ数の割合が 0% である点（図中のオレンジの点）は、集計データのみから推定する CFDM に相当する。提案手法とマルコフモデルについて、どのホールにおいても移動履歴の割合が大きいほど誤差が小さくなった。

各手法における推定精度の検証を行う。CFDM の誤差よりも、すべての割合における提案手法の誤差が小さいことから、CFDM よりも提案手法の推定精度は上回っているといえる。また、提案手法とマルコフモデルについて、どのホールにおいても右下がりの結果となったことから、個人の移動履歴の量が増やすほど推定精度が良くなるといえる。個人の移動履歴の中に真の移動情報が含まれているため、これは当然といえる結果である。着目すべきところは、個人ユーザ数の割合が少ないときの推定精度である。割合が大きくなっていくにつれて、誤差の減少は緩やかになっていることから、割合が少ないときの推定精度の改善が大きい。また、個人ユーザ数の割合が 10% 以下のとき、どのホールにおいても提案手法の推定精度が上回っている。つまり、個人の移動履歴の量が少なきときは、集計データも併せて推定する本手法が有効であるといえる。したがって、現実的なデータ取得の背景における、本手法の有効性を示すことができた。

6.4 定性評価

定性評価として、イベント会場におけるブース間の移動を、矢印を用いて可視化した。 θ_{ij} の大きさに応じて矢印の太さを定めて、移動確率が大きいブース間の矢印は太くした。つまり、ブース間の移動の頻度が矢印の太さで分かるようになっている。可視化の対象はブース数の多いホール 1 とブース数の少ないホール 3 とした。個人ユーザ数の割合が 6% のときの提案手法とマルコフモデル、CFDM における推定を行う。各手法で推定した移動確率 θ^{est} と、真の移動確率 θ^{true} について、ホール 1 とホール 3 の結果をそれぞれ図 2, 3 に示す。全ブース間の移動確率を図示すると煩雑になってしまうため、確率が低いブース間の移動は除外してある。

CFDM (図 2a, 3a) は、真の移動確率や他の手法と比べて、不要な矢印が多く可視化されてしまったことが分かる。つまり、集計データのみから推定すると、移動における解を選別できずに実際に存在しない人流が推定されてしまう。対して、マルコフモデル (図 2b, 3b) は、真の移動確率や他の手法と比べて矢印の数が少ないうえに太さが偏っている。つまり、偏った移動傾向を示している。したがって、個人の移動履歴の割合が少ないときには、真の移動確率 (図 2d, 3d) である全体の人流を把握することが難しいといえる。これらの手法に比べて、提案手法 (図 2c, 3c) は両者の課題点を解決した図示ができています。CFDM における不要な矢印を除去しつつ、マルコフモデルにおける移動傾向の偏りを緩和することによって、真の移動確率 (図 2d, 3d) に非常に近い推定を実現した。以上のことから、移動の可視化における提案手法の有用性を示すことができた。

この図により、一目で頻繁な移動を把握することができるようになっている。あるブースに滞在しているユーザに対して、矢印の太い移動先のブースは、次に向かう可能性が高いといえる。応用先として、ユーザに対して今後のイベント情報や混雑状況といったブース情報の提示などが考えられる。ユーザの行く予定のブース情報を事前に提示することができれば、イベント会場における移動の効率化が期待できる。

7 おわりに

本論文では、少量の移動履歴と大規模な集計データから、人流を高精度に推定する手法を提案した。共通の移動確率に従って移動するという仮説に基づいた確率モデルを学習した。実験では、実際のイベント会場におけるデータを用いて、提案手法の有効性を検証した。従来手法と比べて、少量の移動履歴を用いたときの提案手法について、推定精度の向上や可視化の結果が優れていることを示した。これにより、提案手法がデータ取得の現状を考慮した有効な手法であることが確認できた。

今後の課題としては、推定精度のさらなる向上が挙げられる。今回推定した移動確率 θ_{ij} は、時刻に依存せず場所のみに依存する変数とした。しかし、イベントにおける盛り上がりやショッピングモールにおける特売など、時間帯によって移動の傾向は変化する可能性が高い。VCGMM [9] のように、時間帯で変化する移動確率とすることが考えられる。他にも、移動制約に関するハイパーパラメータ λ の見直しや、集計データと個人の移動履歴に関する重み付けなどが挙げられる。また、観測点間の移動時間を同時推定することによる性能向上 [22]、パラメータ因子分解による効率的な推定 [2]、ニューラルネットを用いた観測点間の移動確率の非線形モデリング [8] のような拡張が考えられる。

【文献】

- [1] 位置情報の利用に対する意識調査, 2014. <http://www.soumu.go.jp/iicp/chousakenkyu/data/research/survey/telecom/2014/location-info.pdf>.
- [2] Y. Akagi, T. Nishimura, T. Kurashima, and H. Toda. A fast and accurate method for estimating people flow from spatiotemporal population data. In *Proceedings*

of the Twenty-Seventh International Joint Conference on Artificial Intelligence, pages 3293–3300, July 2018.

- [3] R. H. Byrd, P. Lu, J. Nocedal, and C. Zhu. A limited memory algorithm for bound constrained optimization. *SIAM J. Sci. Comput.*, 16(5):1190–1208, Sept. 1995.
- [4] H. Cheng, J. Ye, and Z. Zhu. What's your next move: User activity prediction in location-based social networks. *SIAM International Conference on Data Mining*, pages 171–179, 2013.
- [5] F. Evennou and F. Marx. Advanced integration of wifi and inertial navigation systems for indoor mobile positioning. *EURASIP J. Appl. Signal Process.*, 2006:164–164, Jan. 2006.
- [6] Y. Ge, H. Xiong, A. Tuzhilin, K. Xiao, M. Gruteser, and M. Pazzani. An energy-efficient mobile recommender system. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 899–908, 2010.
- [7] A. Goldfarb and C. Tucker. Online display advertising: Targeting and obtrusiveness. *Marketing Science*, 30(3):389–404, May 2011.
- [8] T. Iwata and H. Shimizu. Neural collective graphical models for estimating spatio-temporal population flow from aggregated data. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence*, to appear.
- [9] T. Iwata, H. Shimizu, F. Naya, and N. Ueda. Estimating people flow from spatiotemporal population data via collective graphical mixture models. *ACM Trans. Spatial Algorithms Syst.*, 3(1):2:1–2:18, May 2017.
- [10] J. Krumm. A markov model for driver turn prediction. *SAE 2008 World Congress*, April 2008.
- [11] A. Kumar, D. Sheldon, and B. Srivastava. Collective diffusion over networks: Models and inference. *CoRR*, abs/1309.6841, 2013.
- [12] T. Kurashima, T. Iwata, G. Irie, and K. Fujimura. Travel route recommendation using geotags in photo sharing sites. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, pages 579–588, 2010.
- [13] L.-P. Liu, D. Sheldon, and T. G. Dietterich. Gaussian approximation of collective graphical models. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, pages II–1602–II–1610. JMLR.org, 2014.
- [14] T. Nguyen, A. Kumar, H. C. Lau, and D. Sheldon. Approximate inference using dc programming for collective graphical models. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51, pages 685–693, 09–11 May 2016.
- [15] I. Okajima, S. Tanaka, M. Terada, D. Ikeda, and T. Nagata. Supporting growth in society and industry using statistical data from mobile terminal networks — overview of mobile spatial statistics —. Technical Report 3, NTT DOCOMO Technical Journal, 2012.
- [16] F. Palumbo, P. Barsocchi, S. Chessa, and J. C. Augusto. A stigmergic approach to indoor localization using bluetooth low energy beacons. In *2015 12th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6, Aug 2015.
- [17] D. Sheldon, T. Sun, A. Kumar, and T. Dietterich. Approximate inference in collective graphical models. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28, pages 1004–1012, 17–19 Jun 2013.

- [18] D. R. Sheldon and T. G. Dietterich. Collective graphical models. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 1161–1169. Curran Associates, Inc., 2011.
- [19] R. Simmons, B. Browning, Y. Zhang, and V. Sadekar. Learning to predict driver route and destination intent. In *2006 IEEE Intelligent Transportation Systems Conference*, pages 127–132, Sept 2006.
- [20] X. Song, Q. Zhang, Y. Sekimoto, and R. Shibasaki. Prediction of human emergency behavior and their mobility following large-scale disaster. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 5–14, 2014.
- [21] T. Sun, D. Sheldon, and A. Kumar. Message passing for collective graphical models. In *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37*, pages 853–861, 2015.
- [22] Y. Tanaka, T. Iwata, T. Kurashima, H. Toda, and N. Ueda. Estimating latent people flow without tracking individuals. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, pages 3556–3563, July 2018.

河井 悠佑 Yusuke KAWAI

名古屋大学大学院情報学研究科博士前期課程在学中。2017 年名古屋大学工学部電気電子・情報工学科卒業。日本データベース学会学生会員。

田中 佑典 Yusuke TANAKA

日本電信電話株式会社 サービスエボリューション研究所 研究員。京都大学情報学研究科システム科学専攻 博士課程在学中。2011 年神戸大学工学部電気電子工学科卒業。2013 年京都大学情報学研究科システム科学専攻修士課程修了。同年、日本電信電話株式会社入社。以来、機械学習、データマイニング等の研究に従事。情報処理学会、電子情報通信学会、人工知能学会、各会員。

戸田 浩之 Hiroyuki TODA

日本電信電話株式会社 NTT サービスエボリューション研究所主幹研究員。1997 年名古屋大学工学部材料プロセス工学科卒業。1999 年同大学院工学研究科材料プロセス工学専攻博士課程前期課程修了。同年、日本電信電話株式会社入社。以来、情報検索、データマイニングに関する研究開発に従事。2007 年筑波大学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後期課程修了。博士（工学）。ACM, 情報処理学会、電子情報通信学会、人工知能学会、日本データベース学会、各会員。

石川 佳治 Yoshiharu ISHIKAWA

名古屋大学大学院情報学研究科教授。1989 年 筑波大学第三学群情報学類卒。1994 年 同大学大学院博士課程工学研究科単位取得退学。同年 奈良先端科学技術大学院大学助手。1999 年 筑波大学電子・情報工学系講師。2004 年 同助教授。2006 年 名古屋大学情報連携基盤センター教授。2013 年 現職に至る。博士（工学）（筑波大学）。データベース、データ工学に興味をもつ。ACM, IEEE CS, 情報処理学会、電子情報通信学会、人工知能学会、日本データベース学会、各会員。