# 携帯電話通信時に得られる疎な位置情報履歴を用いた有意位置検出

黒川　茂莉[†a)]　横山　浩之[†]　吉井　和佳[††]　麻生　英樹[††]

# Discovery of Personally Meaningful Places Using Sparse Location Information Associated with Telecommunication Histories of Mobile Phones

Mori KUROKAWA[†a)], Hiroyuki YOKOYAMA[†], Kazuyoshi YOSHII[††], and Hideki ASOH[††]

**あらまし**　本論文では，携帯電話による通信時に取得される空間的粒度が粗く，時間間隔も一定ではない疎な位置情報履歴から個人の有意位置を検出するための手法を提案する．有意位置とは，自宅や職場，行きつけの店がある町など，ある程度の滞在頻度がある場所を指す．有意位置においては，空間的・時間的に近い範囲内で通信がまとまって行われると考えられるため，位置情報履歴をクラスタリングすることで有意位置を検出することができる．しかし，位置情報の空間的な近接性に着目した従来のクラスタリング手法では，移動中に発生する通信の影響で検出精度が悪化するという問題があった．この問題を解決するため，本研究では位置情報の時間的な近接性に着目し，通信基地局を単語，ある一定の時間窓に含まれる位置情報履歴（通信基地局群）を文書とみなすことで，文書集合の潜在的トピックを推定する手法である潜在的ディリクレ配分法（LDA）を適用する．12名の通信履歴を用いた実験の結果，従来手法よりも優れた有意位置検出精度を達成できることが分かった．

**キーワード**　有意位置，携帯電話，通信履歴，クラスタリング，潜在的ディリクレ配分法（LDA）

## 1. まえがき

携帯電話は利用者とほとんど常に行動を共にする電子デバイスであるので，日常生活の様々な記録をライフログとして残すための手軽で便利な手段であると考えられている [1]．近年，Twitter (注1) や Google Buzz (注2) などに代表されるように，家の内外問わずに気軽に自らの身の回りの情報を発信できる（ポストする）ような Web サービスが人気を博している．これらのサービスにおいては，各ポストに GPS 情報（全地球測位システム：GPS）を付加することができ，あとから読み返してみたときに，自らの記憶と関連づけるための重要な手掛りとなっている．しかし，Web サービス上には利用者が手動でポストを行ったときのみに位置情報が記録されるので，記録できる総量には限界がある．また，携帯電話の GPS 機能を常時有効にして携帯端末内に位置情報を蓄積していく方法も考えられるが，バッテリーの消耗が早まる問題がある．

我々は，利用者に負担をかけることなく日常活動の記録を残す手段として，通信事業者のインフラで取得できる通信記録を用いる方法に着目している．ネットワーク上の膨大な通信は記録され，インフラの異常や障害を発見し，原因を追究して迅速に復旧するために利用されている．また，認証や課金を行う上でも，通信に関わる記録が必要である．通常，こうした記録は特定の目的のために必要な期間だけ保管され，期限が過ぎれば消去されてきた．しかし，これをライフログ分析に利用できれば，利用者に対して有用な情報をフィードバックすることができる．

事業者のインフラで収集される通信記録は，全体の総量は巨大であるものの，個人の通信記録のみに着目すると空間的・時間的に疎である点で特徴的である．携帯電話を用いた通信（通話やメールの送受信）が発生したときのみに限り，その時刻と接続先基地局（位

† （株）KDDI 研究所，ふじみ野市
KDDI R&D Laboratories Inc., 2–1–15 Ohara, Fujimino-shi, 365–8502 Japan
†† 独立行政法人産業技術総合研究所，つくば市
National Institute of Advanced Industrial Science and Technology, AIST Tsukuba Central 2, 1–1–1 Umezono, Tsukuba-shi, 305–8568 Japan
a) E-mail: mo-kurokawa@kddilabs.jp

(注1)：http://twitter.com/
(注2)：http://www.google.com/buzz

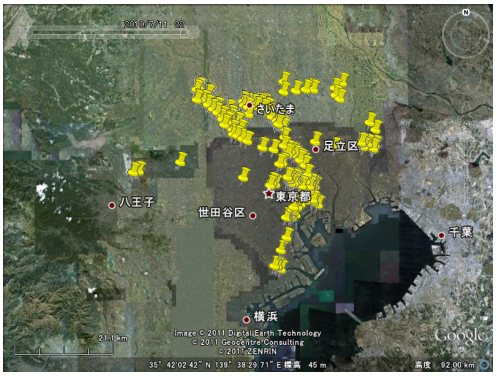



図 1　ある被験者の 1 か月間の位置情報履歴
Fig. 1　Location history in one month.

置情報）とが記録されている．しかし，GPS データとは異なり，空間的な粒度は基地局の立地密度に依存するので，地域によっては疎になりやすい．また，時間間隔が一定ではなく，利用者によっては，通信は朝晩に集中していて昼間はほとんど発生しないという場合もある．例として，ある利用者の 1 か月間の位置情報履歴を図 1 に示す．図中のマーカはこの間に行われた通信の接続先基地局の位置を示している．

本研究では，こうした特徴をもつ位置情報履歴を用いて，利用者にとって意味のある場所（有意位置）[2], [3] を検出することに取り組む．有意位置とは，自宅や職場のように定期的に滞在する場所や，休日に行きつけのお店がある町など，ある程度の滞在時間がある場所を指す．解析結果は，行動パターンが類似した利用者をターゲットにした情報配信などに利用できると考えられる．一般的に，有意位置では空間的・時間的に近い範囲内で通信がまとまって行われると考えられるので，位置情報履歴をいくつかのグループにクラスタリングすることで有意位置を検出することができる．

本論文の構成は以下のとおりである．まず，**2.** で本研究の位置付けを述べる．次に，**3.** で提案手法について説明し，**4.** で評価実験について報告する．最後に，**5.** でまとめと今後の課題を述べる．

## 2.　研究の位置付け

位置情報履歴から有意位置検出を行う目的で，これまで，位置情報の空間的・時間的な情報に着目したクラスタリング手法が考案されてきた．

空間的な情報に着目したクラスタリング手法として，定期的な位置測位や GPS による空間的に高精度かつ時間的に高密度なデータに対して，DBSCAN と呼ばれる位置情報の局所性に基づく凝集型のクラスタリング方法 [4] や二次元平面上における無限混合ガウスモデルを用いた方法 [5] が提案されている．また，通信に利用された基地局の空間的に疎な位置情報履歴に対しても，Leader Algorithm と呼ばれる凝集型のクラスタリング方法 [6] が提案されている．しかし，このような手法では，有意位置間を電車などで移動中に発生する通信（有意位置における通信ではない）の影響を受けて，推定された有意位置の位置がずれてしまう問題があることが知られている [7]．

一方，時間情報を利用したクラスタリング手法として，滞在時間に基づく方法 [7], [8] や滞在地における基地局の切り換わり回数に基づく方法 [9] が提案されている．[7] では，位置情報を時間順に逐次的に処理し，二点間の距離があるしきい値以下のもの同士を逐次的に併合した上で，滞在時間が一定のしきい値以上のクラスタを選択する手法を提案している．[8] では，$k$-means 法を改良したクラスタリングの結果に対し，総滞在時間の基準でフィルタリングする手法を提案している．[9] では，有意位置での通信に使用する基地局数は少数に限られるという仮定のもと，通信に利用された基地局の履歴に対して，基地局の切り換わり回数の上限値（例えば 3 回）を設ける手法を提案している．

本研究では，位置情報の空間的・時間的な局所性に着目したクラスタリング手法を提案する．通信時に得られる位置情報履歴は時間的・空間的に疎であるが，あるタイミングでまとまって発生するというバースト性をもつことが知られている [10]．そこで，ある比較的短い時間区間で利用される通信基地局は限られているという性質を利用して，有意位置推定を行うことができる．このとき，位置情報履歴の時間的な密度には個人差があり，基地局の空間的な密度には地域差があるため，有意位置の個数や各有意位置において利用される基地局の個数などを事前に設定することなしに，高精度なクラスタリングを行うことを試みる．

## 3.　有意位置推定手法

本研究では，利用者の行動の背後には有意位置 1，有意位置 2，· · · のような潜在的なクラスがあり，それぞれの潜在的なクラスの中で利用される基地局は偏っている（いくつかの限られた基地局のみが使用される）と仮定する．位置情報履歴に対してある一定の長さをもつ時間窓をシフトさせながら適用し，その中で利用

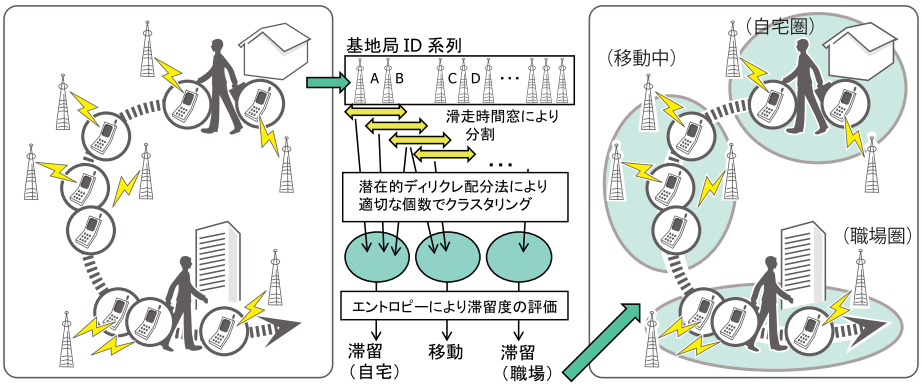



図 2　有意位置推定手法：位置情報履歴を滑走時間窓により分割して LDA を適用
Fig. 2　A proposed method of discovering personally meaningful places: LDA is applied to temrapol segments of the location history.

された基地局の分布（ヒストグラム）を計算すると，有意位置に対応する時間区間ではある特定のクラスのみが出現しやすく，移動中に対応する時間区間では複数のクラスが混ざって出現しやすいなどといったマイニングが可能になると考えられる．

このような潜在的なクラスを発見するため，本研究では潜在的ディリクレ配分法（Latent Dirichlet Allocation: LDA）[11] を利用する．LDA は，もとは自然言語処理分野で提案された文書のためのトピックモデルである．近年，画像処理 [12]，音楽情報処理 [13]，ソーシャルネットワーク解析 [14] などの様々な分野において，物体のカテゴリー，楽曲のジャンル，コミュニティといったデータに内在する潜在的なクラス構造（便宜的に「トピック」と呼ばれる）を推定するために広く利用され，その有効性が確かめられている．更に，検出すべき有意位置の数を個人ごとに適切に推定するため，ノンパラメトリックベイズ理論を用いる．具体的には，LDA をノンパラメトリックベイズ拡張した階層ディリクレ過程 LDA（Hierarchical Dirichlet Process LDA: HDP-LDA）[15] を利用する．

### 3.1　位置情報履歴に対する LDA の適用

LDA を位置情報履歴に対して適用するには，「単語」及び「文書」に相当するものを定義する必要がある．本研究では，ある短い時間区間において（特に有意位置において）特定の基地局が利用されやすいという性質に着目している．これは，各文書において特定の単語が頻出しやすいという性質と同様である．したがって，基地局を単語，ある一定の時間内に利用された基地局の集合を文書をみなすことで，LDA を適用することができる．ただし，同じ基地局が複数回利用された場合は，それらは別々の観測（同じ単語が複数文書に含まれる）として取り扱うものとする．具体的には，本研究では，時間幅 $T$，シフト幅 $S$ の滑走時間窓を用いて，位置情報履歴から一定幅の時間区間を次々に切り出すことで，時間局所的な位置情報履歴を一つの文書とみなし（図 2），各区間について，自然言語処理分野で Bag-of-Words ベクトルに相当するものとして，その間に通信を行った基地局 ID のヒストグラムを作成する．本研究では，$S$ は 15 分とし，$T$ は 30 分から 150 分まで変化させた．

このように時分割された基地局の集合に対して LDA を適用すると，各時間区間（文書）$d$ におけるトピックの混合比 $\boldsymbol{\theta}_d$ 及び各トピック $k$ における基地局（単語）の使用頻度分布 $\boldsymbol{\phi}_k$ が得られる．本研究では，ある有意位置において利用される基地局の共起パターンは，あるトピックにおける基地局の分布と対応していると仮定し，$\boldsymbol{\phi}_k$ を観察することで有意位置を推定する．各トピック $k$ に対応する地理空間上の位置を求めるには，基地局の分布 $\boldsymbol{\phi}_k$ のもとでの基地局の地理空間上の位置の重み付き平均を計算した．

しかし，実際には，移動中に相当するトピックも形成されてしまう問題がある．この問題を解決するには，有意位置では少数の基地局と通信を行う可能性が高く，移動中には多くの基地局と通信を行う傾向があることに着目する．すなわち，各トピック $k$ のエントロピー $H(\boldsymbol{\phi}_k)$ の逆数を計算することで，有意位置らしさを評価することができる．

### 3.2　潜在的ディリクレ配分法

本節では，基本的な LDA の概要と，確率モデルの定式化及び学習方法について説明する．

**3.2.1** 各文書に対するトピックの混合

LDA は，Blei らによって提案された文書集合に対するベイズ的な生成モデルである [11]．今，全データ中の文書数を $D$，単語の語彙の集合を $\boldsymbol{W}$，そのサイズを $V$ とする．各文書 $d \in \{1,\cdots,D\}$ は，$K$ 個の「トピック」から構成されていると仮定し，その混合比を $K$ 次元ベクトル $\boldsymbol{\theta}_d$ で表す．ここで，トピックとは，語彙の集合 $\boldsymbol{W}$ に含まれる各単語が出現する確率の集合を意味し，各トピック $k \in \{1,\cdots,K\}$ は $V$ 次元ベクトル $\boldsymbol{\phi}_k$ で表すことができる．ただし，$\boldsymbol{\theta}_d$ 及び $\boldsymbol{\phi}_k$ のいずれに関しても，各要素は確率値であり，全ての要素を足し合わせると 1 になるよう正規化されている．通常の混合モデルでは各文書を $K$ 個のうちのいずれかのトピックに排他的に割り当てるのに対し，LDA では，文書（時間窓）内の各単語（基地局）をそれぞれ異なるトピックに排他的に割り当てる．したがって，各文書が複数のトピックの混合から生成される点が特徴的である．なお，トピックとは必ずしも人間の感覚と合致したものではなく，文書データから自己組織的に学習されるものであることに注意する．

**3.2.2** 確率モデルの定式化

本項では，LDA の確率モデルの定式化について説明する．各文書 $d$ に含まれる単語数を $N_d$ とし，文書 $d$ を観測変数の集合 $\boldsymbol{X}_d = \{x_{d,1},\cdots,x_{d,N_d}\}$ で表す．文書 $d$ を複数のトピックから構成されているように表現するには，文書 $d$ に含まれる各単語 $x_{d,n} \in W$ $(n = \{1,\cdots,N_d\})$ が別々のトピックに所属することを許容することが必要である．すなわち，各観測変数について，それが $K$ 個のうちどのトピックに所属するかを示す潜在変数の集合 $\boldsymbol{Z}_d = \{z_{d,1},\cdots,z_{d,N_d}\}$ が存在する．ここで，$z_{d,n} \in \{1,\cdots,K\}$ であり，その値は文書 $d$ におけるトピックの混合比 $\boldsymbol{\theta}_d$ に従って確率的に決まる．ここで，$z_{d,n} = k$ と決まれば，単語分布 $\boldsymbol{\phi}_k$ に従って単語 $x_{d,n}$ が確率的に決まる．

以上の確率モデルを構成する確率変数は以下のとおりである．まず，全文書における観測変数及び潜在変数をまとめて，$\boldsymbol{X} = \{\boldsymbol{X}_1,\cdots,\boldsymbol{X}_D\}$ 及び $\boldsymbol{Z} = \{\boldsymbol{Z}_1,\cdots,\boldsymbol{Z}_D\}$ としておく．また，パラメータ（各文書におけるトピックの混合比・各トピックにおける単語分布）に関しても，$\boldsymbol{\theta} = \{\boldsymbol{\theta}_1,\cdots,\boldsymbol{\theta}_D\}$ 及び $\boldsymbol{\phi} = \{\boldsymbol{\phi}_1,\cdots,\boldsymbol{\phi}_K\}$ としておく．

今，各文書 $d$ におけるトピックの混合比 $\boldsymbol{\theta}_d$ 及び各トピック $k$ における単語分布 $\boldsymbol{\phi}_k$ も未知であるから，それらの不確実性を適切に取り扱いたい．そこで，これら離散分布に対する共役事前分布としてディリクレ分布（Dir と表記）を導入する．まとめると，LDA の生成モデル（図 3 (a)）は以下のようになる．

(1) 各トピック $k$ $(k = 1,\cdots,K)$ に対して
  (1.1) $\mathrm{Dir}(\beta\boldsymbol{\tau})$ から単語分布 $\boldsymbol{\phi}_k$ を生成
(2) 各文書 $d$ $(d = 1,\cdots,D)$ に対して
  (2.1) $\mathrm{Dir}(\alpha\boldsymbol{\pi})$ からトピックの混合比 $\boldsymbol{\theta}_d$ を生成
  (2.2) 各単語 $n$ $(n = 1,\cdots,N_d)$ に対して
    (2.2.1) 離散分布 $\boldsymbol{\theta}_d$ から潜在変数 $z_{d,n}$ を生成
    (2.2.2) 離散分布 $\boldsymbol{\phi}_{z_{d,n}}$ から観測変数 $x_{d,n}$ を生成

ここで，$\alpha$，$\boldsymbol{\pi}$，$\beta$，$\boldsymbol{\tau}$ は超パラメータであり，$\alpha$ 及び $\beta$ はディリクレ分布の集中度，$\boldsymbol{\pi}$ 及び $\boldsymbol{\tau}$ はディリクレ分布の平均である．$\boldsymbol{\pi}$ 及び $\boldsymbol{\tau}$ は，全要素を足して 1 になるよう正規化されている（通常は一様分布）．

**3.2.3** 確率モデルの学習

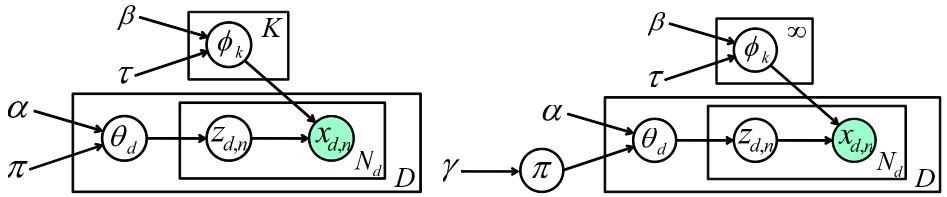
確率モデルの学習とは，文書データが与えられたときに，潜在変数 $z_{d,n}$ 及び未知のパラメータ $\boldsymbol{\phi}_k$ 及び $\boldsymbol{\theta}_d$ の事後分布を推定することである．しかし，真の事後分布を解析的に計算することはできないため，何らかの近似解法が必要になる．Blei らは，変分ベイズ法（Variational Bayes; VB）を用いて，真の事後分布を解析的に導出可能な分布で近似する学習法を提案している [11]．一方，Griffiths らは，周辺化ギブスサンプリング（Collaped Gibbs Sampling; CGS）を用いて，パラメータである $\boldsymbol{\phi}_k$ 及び $\boldsymbol{\theta}_d$ を積分消去した空間で，潜在変数 $z_{d,n}$ の値を効率的にサンプリングしていく方法を提案している [16]．Teh らは，CGS と同様に，パラメータを積分消去した空間で変分ベイズ法により潜在変数 $z_{d,n}$ の事後分布を推定する周辺化変分ベイズ法（Collaped Variational Bayes; CVB）を提案している [17]．このように，パラメータを周辺化して推論を行う場合でも，$\boldsymbol{\phi}_k$ 及び $\boldsymbol{\theta}_d$ の事後分布や期待値を求めることは容易である．

### 3.3 階層ディリクレ過程潜在的ディリクレ配分法

次に，HDP-LDA の概要と，確率モデルの定式化及び学習方法について説明する．ノンパラメトリックベイズモデルの解説としては，[18] も参照されたい．

**3.3.1** トピック数の自動調節

これまでトピック数 $K$ は既知であるとしてきたが，文書データに合わせて適切な値を設定するのは容易ではない．この問題を解決するため，文書データに合わせて自動的にトピック数を調節することができるよう LDA を拡張したものが HDP-LDA である．HDP-LDA は，LDA におけるトピック数 $K$ を無限に発散




(a) LDA　(b) HDP-LDA

図 3　LDA と HDP-LDA の確率的生成モデルのグラフィカル表現
Fig. 3　Graphical representations of LDA and HDP-LDA.

させたときの極限に等しい．もし，文書データが無限に存在すれば，その生成過程を記述するのに無限個のトピックが必要になるであろう．しかし，現実的には文書データは有限であるので，本来無限個存在するうちでたかだか有限個のトピックを選んで利用するだけで十分である．すなわち，文書データの複雑さに合わせて，「実質的な」トピック数が自動的に調節される．

**3.3.2**　確率モデルの定式化

本項では，HDP-LDA の確率モデルの定式化について説明する．いま，LDA において $K \to \infty$ の極限を考えると，文書 $d$ におけるトピックの混合比 $\boldsymbol{\theta}_d$ は無限次元の離散分布になる．したがって，トピックの混合比に対するディリクレ分布 $\mathrm{Dir}(\alpha\boldsymbol{\pi})$ も無限次元となる．実は，この無限次元のディリクレ分布は，$\alpha$ を集中度，$\boldsymbol{\pi}$ を基底測度とするディリクレ過程（Dirichlet Process: DP）と等価であり，$\mathrm{DP}(\alpha, \boldsymbol{\pi})$ で表す．DP から生成されるトピックの混合比の期待値は $\alpha$ にかかわらず $\boldsymbol{\pi}$ であるが，$\alpha$ が大きいほど $\boldsymbol{\pi}$ と似通った混合比 $\boldsymbol{\theta}_d$ がサンプルされる確率が高くなる．すなわち，DP を用いると，無限次元の離散分布 $\boldsymbol{\pi}$（基底測度）に基づき，それとは少し異なる無限次元の離散分布 $\boldsymbol{\theta}_d$ を確率的に生成することができる．

無限次元の離散分布 $\boldsymbol{\pi}$ をどのように生成するかという問題に対しては，階層ディリクレ過程（Hierarchical Dirichlet Process: HDP）[15] を構成することで対処可能である．すなわち，DP の基底測度 $\boldsymbol{\pi}$ に対する超事前分布として，DP の実現方法である棒折り過程（Stick-Breaking Process: SBP）[19] を設定することができる．すなわち，上位階層の DP の集中度を $\gamma$ とすると，$\mathrm{SBP}(\gamma)$ から $\boldsymbol{\pi}$ が生成されると考える．

これらをまとめると，HDP-LDA の生成モデルは，**3.2.2** で述べた LDA の生成モデルに少し変更を加えることで，以下のとおり構成できる（図 3 (b)）．

(1) 各トピック $k$ $(k = 1, \cdots, \infty)$ に対して
　(1.1) $\mathrm{Dir}(\beta\boldsymbol{\tau})$ から単語分布 $\boldsymbol{\phi}_k$ を生成
(2) $\mathrm{SBP}(\gamma)$ からトピックの混合比 $\boldsymbol{\pi}$ を生成
(3) 各文書 $d$ $(d = 1, \cdots, D)$ に対して
　(3.1) $\mathrm{DP}(\alpha, \boldsymbol{\pi})$ からトピックの混合比 $\boldsymbol{\theta}_d$ を生成
　(3.2) 各単語 $n$ $(n = 1, \cdots, N_d)$ に対して
　　(3.2.1) 離散分布 $\boldsymbol{\theta}_d$ から潜在変数 $z_{d,n}$ を生成
　　(3.2.2) 離散分布 $\boldsymbol{\phi}_{z_{d,n}}$ から観測変数 $x_{d,n}$ を生成

ここで，超パラメータである DP の集中度 $\alpha$，$\gamma$ も未知であるので，ほぼ無情報となるようなガンマ分布を事前分布として与える．$\boldsymbol{\tau}$ は一様分布とする．

**3.3.3**　確率モデルのベイズ学習

ベイズ学習の目的は，LDA と同様に，文書データ $\boldsymbol{X}$ が与えられたときに，潜在変数 $\boldsymbol{Z}$ 及び未知のパラメータ $\boldsymbol{\theta}$，$\boldsymbol{\phi}$，$\boldsymbol{\pi}$，$\alpha$，$\gamma$ の事後分布 $p(\boldsymbol{Z}, \boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\pi}, \alpha, \gamma | \boldsymbol{X})$ を推定することである．HDP-LDA においても，周辺化変分ベイズ法 [20] あるいは周辺化ギブスサンプリング [15] を用いて近似推定を行うことができる．これらはいずれも反復的な解法であるが，一般に，周辺化変分ベイズ法の方が収束が早く，周辺化ギブスサンプリングの方が精度が高い傾向がある．

求まった各文書 $d$ のトピックの混合比 $\boldsymbol{\theta}_d$ は無限次元であるが，ある限られた要素のみがある程度の大きさをもち，他は全てほぼゼロであるようなベクトルとなっている．そのため，有限の大きさの文書データでは，ある限られた個数のトピックしか出現しない．

## 4. 評価実験

本章では，提案手法の有効性を検証するために行った評価実験について報告する．

### 4.1 実験用データ

本研究では，実際の携帯電話の通信時に利用された基地局の位置情報履歴データを対象として実験を行った．被験者は首都圏在住の会社員計 12 名であり，普段どおり日常生活を送ってもらい，通信履歴を最大 4 週間分取得した．履歴取得後，データ収集期間中の各被験者の滞留地点（自宅，職場，及び各週の出かけ先）

表 1　収集したデータの概要
Table 1　Summary of collected data.

| 被験者 No. | 平均レコード数（1 日当り） | 総通信基地局数 | 総滞留地点数 |
|---|---|---|---|
| 1 | 7.00 | 24 | 5 |
| 2 | 41.75 | 107 | 8 |
| 3 | 17.18 | 36 | 9 |
| 4 | 42.89 | 103 | 11 |
| 5 | 54.82 | 127 | 12 |
| 6 | 30.82 | 65 | 8 |
| 7 | 35.81 | 82 | 9 |
| 8 | 10.48 | 57 | 12 |
| 9 | 43.25 | 54 | 6 |
| 10 | 13.00 | 37 | 9 |
| 11 | 40.32 | 63 | 7 |
| 12 | 114.25 | 75 | 7 |

についてのアンケート調査も行った．

表 1 のとおり，被験者ごとの 1 日当りのレコード数（通話，SMS や E メールの送受信，その他のデータ通信などによる基地局との通信数）の平均は，多い人で 100 超，少ない人で 10 未満であった．一般的な通信の利用状況と照らし合わせると，通話の回数は 1 人 1 日当り 1.4 程度 [21]，E メールの送受信の回数は 1 日当り 11〜20 通の人が最も多い [22] というデータがある．以上を足し合わせると 12〜22 回程度の通信数となるが，SMS，Web ブラウジング，アプリケーションによる通信数の分が上乗せされるため，平均的にそれよりも多い通信数となっている．

全期間中に観測された通信基地局数（LDA における単語の語彙サイズ $V$ に相当）は，多い人で 120 超，少ない人で 30 未満であった．アンケートの結果得られた滞留地点数は，5〜12 であった．ここで，滞留地点数は，自宅及び職場の 2 地点に加えて期間を通じて複数回アンケート回答のあったお出かけ先について重複を排除した数を合計した数である．

### 4.2 実験条件

提案する有意位置検出手法として，LDA に基づくものと HDP-LDA に基づくものの 2 種類を適用した．滑走時間窓の時間幅は，$T = 30$ 分，60 分，90 分，120 分，150 分の 5 種類，滑走窓のシフト幅は，$S = 15$ 分で固定とした．LDA 及び HDP-LDA いずれの場合に対しても，周辺化ギブスサンプリング法（CGS）及び周辺化変分ベイズ法（CVB）を用いて学習を行った．それぞれの組合せについて，CGS-LDA [16]，CVB-LDA [17]，CGS-HDP-LDA [15]，CVB-HDP-LDA [20] と呼ぶことにする．学習時の反復回数は，CVB が 100 回，CGS が 2000 回とした．基本的に，事前分布はほぼ無情報事前分布となるように与えた．ただし，トピックの単語分布 $\boldsymbol{\theta}$ に対するディリクレ分布の超パラメータ $\beta\boldsymbol{\tau}$ は，各要素が 0.1，1.0，10.0 となるときの 3 種類を試した．すなわち，$\beta = V/10, V, 10V$ の 3 種類である．トピック数については，HDP-LDA の場合は，トピック数の上限は 50（実際にはこれより少ない数のトピックが利用される）とし，LDA の場合は，その限界性能を調べるため，トピック数を $K = 1, \cdots, 20$ と変化させて，実験を行った．

比較のために，従来手法として Leader Algorithm [6] と無限混合ガウス分布（Infinite Gaussian Mixture Model: iGMM）[5] とを適用した．これらの手法においては，位置情報に付与されたタイムスタンプは利用されず，位置情報履歴は二次元平面上の順不同なデータの集合として取り扱われる．すなわち，通信が記録されるたびに，利用された基地局の緯度･経度を一つのデータ点として蓄積される．Leader Algorithm の適用にあたっては，文献 [6] と同様に，あらかじめ基地局リストを各基地局と通信した日数の降順に並べ換えた．また，クラスタリングの際の距離のしきい値は 2000 m，3000 m，4000 m，5000 m，10000 m の場合を実験した．iGMM の事後分布の計算には変分ベイズ法（VB）を用い，反復回数は 100 回とした．事前分布はほぼ無情報事前分布とし，混合数の上限は 50 とした．

### 4.3 評価方法

有意位置検出精度を評価するため，適合率と再現率，及びそれらの調和平均として求められる $F$ 値を計算した．適合率と再現率は次式により計算した．

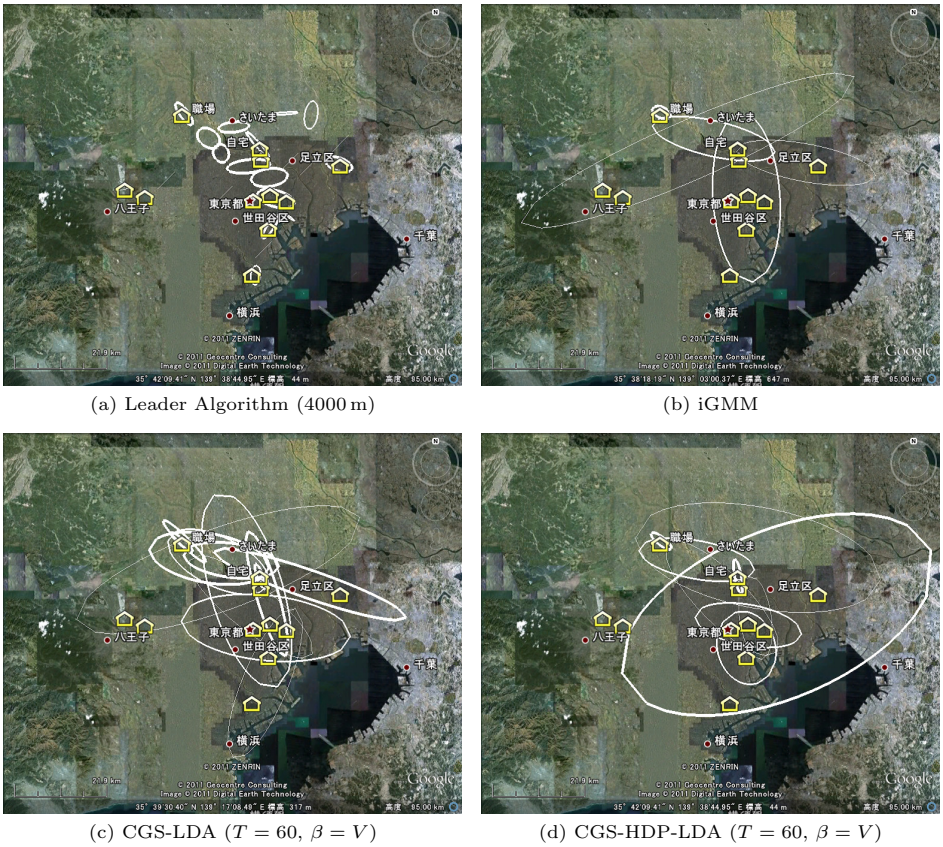
$$\text{適合率} = \frac{\text{アンケートと対応がとれたクラスタ数}}{\text{足切り後のクラスタ数}}$$

$$\text{再現率} = \frac{\text{アンケートと対応がとれたクラスタ数}}{\text{アンケート結果の滞留地点数}}$$

以降，これらの算出方法について具体的に説明する．

まず，各種のクラスタリング手法の適用結果に対し，各クラスタに対して割り当てられた延べ基地局数が，全クラスタに関する合計の 1%以下のものを除外した．各クラスタに対して割り当てられた延べ基地局数 $n_k$ は，各時間区間に含まれる基地局に対して割り当てられたクラスタ（トピック）が何であるかを調べ，クラスタごとにカウントしたものである．

$$n_k = \sum_{d,n} I[z_{d,n} = k]$$




(a) Leader Algorithm (4000 m)　(b) iGMM

(c) CGS-LDA ($T = 60$, $\beta = V$)　(d) CGS-HDP-LDA ($T = 60$, $\beta = V$)

図 4　被験者 No.5 の各手法の有意位置抽出結果の比較
Fig. 4　Personally meaningful place maps for subject No.5 extracted by tested methods.

ここで，$I[\,]$ は $[\,]$ 内の条件を満たす場合に 1，それ以外の場合に 0 をとる指示関数である．

次に，残ったクラスタを滞留度に基づいて並べ換えた．LDA 及び HDP-LDA の場合は，各トピック $k$ に対して，基地局分布 $\boldsymbol{\phi}_k$ のエントロピー $H(\boldsymbol{\phi}_k)$ の逆数に基づいて滞留度を算出した．Leader Algorithm の場合は，各クラスタに対する位置情報の割り当て数，iGMM の場合は，各クラスタ（二次元ガウス分布）の第一主成分の分散の逆数に基づいて滞留度を算出した．

最後に，滞留度の高さを優先順位として，そのクラスタにアンケート結果の滞留地点が対応するか否かを判定した．具体的には，各クラスタに所属する位置情報（緯度・経度）に対して二次元正規分布を当てはめることで信頼度 = 95%の棄却楕円を描き，棄却楕円領域と棄却楕円の中心の周囲 2 km の円領域との和集合領域内にアンケートで得た滞留地点（駅名）の緯度・経度が入った場合に，当該クラスタがアンケートと対応していると判定した．なお，アンケート回答結果の中に対応づけられる場所が複数ある場合は，回答の記載順が早いものに対応づけた．

### 4.4 実験結果

典型的な 1 人の被験者について，Leader Algorithm，iGMM，LDA，HDP-LDA で得られた有意位置検出結果を図 4 (a)–4 (d) に示す．図中の家のアイコンはアンケート結果の滞留地点を表し，楕円は検出された各クラスタに対して描いた 95%棄却楕円である．太い線で描かれたクラスタほど，滞留度が高いことを表している．これを見ると，LDA，HDP-LDA は，自宅・職場をそれぞれ一つの有意位置クラスタとし，自宅・職場以外の滞留地点を含む大小様々なクラスタを生成していることが分かる．それに対し，Leader Algorithm は領域の小さな多数のクラスタが存在し，iGMM は領域の大きな少数のクラスタが存在することが分かる．

**4.4.1** 適合率・再現率・$F$ 値による評価

表 2 に，適合率と再現率及び $F$ 値を用いた評価結果を示す．これらの値は，各手法をそれぞれの被験者の位置情報履歴データに対して適用して得られた結果を，全ての被験者に関して平均したものである．LDA については，被験者ごとに $F$ 値が最大となったトピック数 $K^*$ を用いた結果を示している．滑走時間窓の時間幅 $T = 60$ を用いた LDA の最大の $F$ 値が最も優れた $F$ 値を達成した．次いで，滑走時間窓の時間幅 $T = 60$ を用い，データに基づいてトピック数を自動的に最適化した HDP-LDA が，二番目に優れた $F$ 値を達成した．一方，Leader Algorithm は，距離のしきい値が大きくなるほど，多くの点を凝集したクラスタが生成されやすく全体のクラスタ数が少なくなるため，適合率が上昇し，再現率は低下する傾向がある．しかし，いずれの場合においても $F$ 値において LDA，HDP-LDA に及ばなかった．iGMM は，領域の大きい少数のクラスタが生成されやすいため，適合率は最も高いが，再現率が低くなる問題があり，やはり $F$ 値において LDA，HDP-LDA に及ばなかった．

表 3 に，各手法による有意位置検出の適合率，再現率，$F$ 値の被験者ごとの値を示した．太字は被験者ごとに $F$ 値が最も高いものである．LDA，HDP-LDA はともに，被験者が申告した滞留地点数に近い適切なクラスタ数になっており，過半数の被験者について適合率，再現率がともに高い．それに対し，Leader Algorithm は，検出クラスタ数が多く，適合率が低い被験者が存在する．特に出張が多い被験者である被験者 No.2 や No.5 に対しては，クラスタ数が多くなる傾向があることが分かる．iGMM は，LDA，HDP-LDA と比べて，適合率の分母の値から分かるように検出クラスタ数が少なく，検出されるクラスタも領域の大き

表 2　適合率・再現率・$F$ 値の比較
Table 2　Results of precision, recall and $F$-measure.

| 有意位置検出手法 | 適合率 | 再現率 | $F$ 値 |
|---|---|---|---|
| Leader Algorithm (2000 m) | 0.276 | 0.736 | 0.385 |
| Leader Algorithm (3000 m) | 0.366 | 0.700 | 0.455 |
| Leader Algorithm (4000 m) | 0.417 | 0.594 | 0.462 |
| Leader Algorithm (5000 m) | 0.403 | 0.499 | 0.420 |
| Leader Algorithm (10000 m) | 0.632 | 0.369 | 0.443 |
| iGMM | 0.951 | 0.402 | 0.551 |
| CGS-LDA ($T = 60$, $\beta = V$) | **0.849** | **0.796** | **0.812** |
| CGS-HDP-LDA ($T = 60$, $\beta = V$) | 0.800 | 0.738 | 0.749 |

表 3　被験者ごとの適合率・再現率・$F$ 値の比較
Table 3　Results of precision, recall and $F$-measure for each subject.

| | Leader Algorithm (4000 m) | | | iGMM | | | CGS-LDA ($T = 60$, $\beta = V$) | | | CGS-HDP-LDA ($T = 60$, $\beta = V$) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 被験者 No. | 適合率 | 再現率 | F 値 | 適合率 | 再現率 | F 値 | 適合率 | 再現率 | F 値 | 適合率 | 再現率 | F 値 |
| 1 | 0.600 (3/5) | 0.600 (3/5) | 0.600 | 1.000 (2/2) | 0.400 (2/5) | 0.571 | 0.8 (4/5) | 0.8 (4/5) | 0.8 | 1.000 (4/4) | 0.800 (4/5) | **0.889** |
| 2 | 0.200 (7/35) | 0.875 (7/8) | 0.326 | 1.000 (3/3) | 0.375 (3/8) | 0.545 | 1.000 (7/7) | 0.875 (7/8) | **0.933** | 0.800 (8/10) | 1.000 (8/8) | 0.889 |
| 3 | 0.571 (4/7) | 0.444 (4/9) | 0.500 | 1.000 (2/2) | 0.222 (2/9) | 0.364 | 0.889 (8/9) | 0.889 (8/9) | **0.889** | 0.857 (6/7) | 0.667 (6/9) | 0.750 |
| 4 | 0.353 (6/17) | 0.545 (6/11) | 0.429 | 1.000 (5/5) | 0.455 (5/11) | 0.625 | 0.917 (11/12) | 1 (11/11) | **0.957** | 0.900 (9/10) | 0.818 (9/11) | 0.857 |
| 5 | 0.348 (8/23) | 0.667 (8/12) | 0.457 | 1.000 (5/5) | 0.417 (5/12) | 0.588 | 0.769 (10/13) | 0.833 (10/12) | **0.8** | 0.889 (8/9) | 0.667 (8/12) | 0.762 |
| 6 | 0.500 (5/10) | 0.625 (5/8) | 0.556 | 1.000 (5/5) | 0.625 (5/8) | **0.769** | 0.667 (6/9) | 0.75 (6/8) | 0.706 | 0.800 (4/5) | 0.500 (4/8) | 0.615 |
| 7 | 0.357 (5/14) | 0.556 (5/9) | 0.435 | 1.000 (4/4) | 0.444 (4/9) | 0.615 | 0.889 (8/9) | 0.889 (8/9) | **0.889** | 0.800 (8/10) | 0.889 (8/9) | 0.842 |
| 8 | 0.250 (4/16) | 0.333 (4/12) | 0.286 | 1.000 (2/2) | 0.167 (2/12) | 0.286 | 1.000 (10/10) | 0.833 (10/12) | **0.909** | 1.000 (8/8) | 0.667 (8/12) | 0.800 |
| 9 | 0.500 (5/10) | 0.833 (5/6) | 0.625 | 1.000 (4/4) | 0.667 (4/6) | **0.800** | 0.571 (4/7) | 0.667 (4/6) | 0.615 | 0.625 (5/8) | 0.833 (5/6) | 0.714 |
| 10 | 0.500 (2/4) | 0.222 (2/9) | 0.308 | 1.000 (3/3) | 0.333 (3/9) | 0.500 | 1.000 (4/4) | 0.444 (4/9) | **0.615** | 0.800 (4/5) | 0.444 (4/9) | 0.571 |
| 11 | 0.444 (4/9) | 0.571 (4/7) | 0.500 | 0.667 (2/3) | 0.286 (2/7) | 0.400 | 0.857 (6/7) | 0.857 (6/7) | **0.857** | 0.625 (5/8) | 0.714 (5/7) | 0.667 |
| 12 | 0.375 (6/16) | 0.857 (6/7) | 0.522 | 0.750 (3/4) | 0.429 (3/7) | 0.545 | 0.833 (5/6) | 0.714 (5/7) | **0.769** | 0.500 (6/12) | 0.857 (6/7) | 0.632 |
| 平均 | 0.417 | 0.594 | 0.462 | 0.951 | 0.402 | 0.551 | 0.849 | 0.796 | 0.812 | 0.800 | 0.738 | 0.749 |
| 標準偏差 | 0.124 | 0.201 | 0.111 | 0.115 | 0.145 | 0.151 | 0.133 | 0.142 | 0.117 | 0.151 | 0.161 | 0.110 |

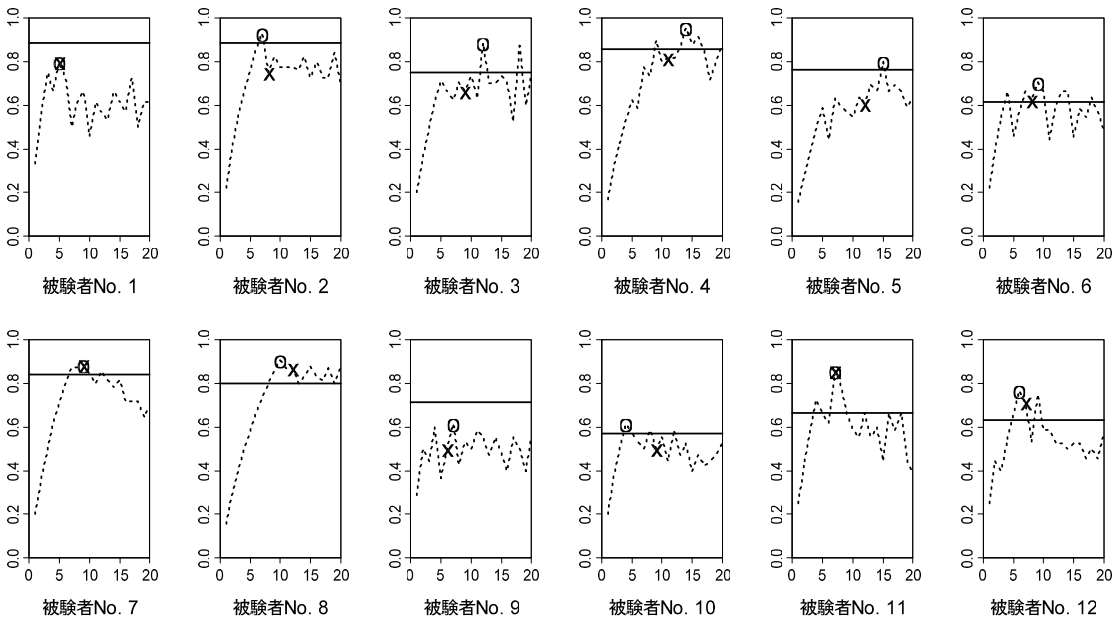



図 5　CGS-LDA（$\beta = V$）のトピック数 $K$ を変化させた場合の被験者ごとの $F$ 値
Fig. 5　Effects of the number of topics $K$.

なクラスタが多いため，適合率は多くの場合で 1.0 となった．適合率が 1.0 で再現率が LDA，HDP-LDA とほぼ対等な被験者 No.6 や No.9 については $F$ 値において LDA，HDP-LDA を上回ったが，その他の被験者については $F$ 値において LDA または HDP-LDA を下回った．

**4.4.2**　トピック数の影響

本項では，LDA のトピック数 $K$ の影響を調査した結果について述べる．図 5 の点線は，CGS-LDA（$\beta = V$）について，被験者ごとにトピック数 $K$ を変化させた場合の $F$ 値を示し，実線は CGS-HDP-LDA（$\beta = V$）の $F$ 値を示している．ここで，滑走時間窓の時間幅は $T = 60$ とした．点 o は，被験者ごとに最適なトピック数 $K^*$ と，そのときの CGS-LDA の $F$ 値を示し，点 x は，アンケートにおいて被験者が滞留地点として回答した有意位置の正解個数 $\bar{K}$ と，そのときの CGS-LDA の $F$ 値を示している．

まず，$K^*$ の場合で比較すると，全体のうち 10 人について，CGS-LDA が $F$ 値において CGS-HDP-LDA を上回ったことが分かる．一方，$\bar{K}$ と $K^*$ は被験者ごとに異なり，3 人の被験者については，$\bar{K}$ と $K^*$ が一致し，残りの被験者については一致しなかった．$\bar{K}$ で比較すると，逆に，全体のうち 7 人について，CGS-HDP-LDA が CGS-LDA を上回ったことが分かる．トピック数 $K$ を変化させた場合，1 から増加させていくにつれて $F$ 値が上昇し，最大値に到達した後は緩やかに低下する傾向がある．低下する幅は，トピック数を事後的に足切りすることによって抑えられている．上下の変動は，$F$ 値が一つの正誤によって変動しやすいことによる．

以上より，最適なトピック数 $K^*$ を選択した場合は，CGS-LDA の方が，$F$ 値において，CGS-HDP-LDA よりよいことが分かる．一方，十分な事前情報が得られないなど，最適なトピック数 $K^*$ を選択することが難しい場合にも，効率的に既存手法を上回る精度が得られるという点で，HDP-LDA も有効性が高いと考える．次項では，HDP-LDA を用いて，データからトピック数を自動的に調節する場合の時間窓幅・学習法・超パラメータの影響について述べる．

**4.4.3**　時間窓幅・学習法・超パラメータの影響

本項では，HDP-LDA を適用する際の，滑走窓の時間幅 $T$，ベイズ学習法の選択（CGS あるいは CVB），単語分布に対するディリクレ事前分布の超パラメータ $\beta$ の影響を調査した結果について述べる．

表 4 は，CGS-HDP-LDA について，適用する滑走窓の時間幅について比較した結果である．滑走窓に関しては，30 分の時間幅の精度は 60 分の時間幅の精度に対して低く，60〜150 分の時間幅の $F$ 値は大きく異

表 6　各手法の有意位置検出結果に表れる自宅，職場の順位の比較
Table 6　Ranking of home and office in extracted personally meaningful places for each algorithm.

| | Leader Algorithm (4000 m) | | iGMM | | CGS-LDA ($T = 60$, $\beta = 1.0$) | | CGS-HDP-LDA ($T = 60$, $\beta = 1.0$) | |
|---|---|---|---|---|---|---|---|---|
| 被験者 No. | 自宅順位 | 職場順位 | 自宅順位 | 職場順位 | 自宅順位 | 職場順位 | 自宅順位 | 職場順位 |
| 1 | 1 | 3 | 2 | 1 | 3 | 1 | 2 | 1 |
| 2 | 1 | 2 | 1 | - | 1 | 2 | 1 | 2 |
| 3 | 2 | - | 1 | - | 1 | 2 | 1 | 2 |
| 4 | 2 | - | 1 | - | 2 | 1 | 2 | 1 |
| 5 | 3 | 7 | 2 | 1 | 3 | 1 | 2 | 1 |
| 6 | 2 | 1 | 2 | 1 | 3 | 1 | 2 | 1 |
| 7 | 3 | 1 | 2 | 1 | 3 | 1 | 2 | 1 |
| 8 | 2 | 1 | 1 | 2 | 4 | 1 | 3 | 1 |
| 9 | 3 | 1 | 2 | 1 | 6 | 1 | 3 | 1 |
| 10 | - | 4 | - | 1 | - | 2 | - | 2 |
| 11 | 7 | 1 | 2 | 1 | 3 | 1 | 2 | 1 |
| 12 | 7 | 1 | 1 | 2 | 1 | 3 | 1 | 3 |

表 4　CGS-HDP-LDA（$\beta = V$）における滑走窓の時間幅 $T$ の適合率・再現率・$F$ 値に対する影響
Table 4　Effects of the time window length $T$.

| 滑走窓の時間幅 $T$ | 適合率 | 再現率 | $F$ 値 |
|---|---|---|---|
| 30 分 | 0.769 | 0.620 | 0.665 |
| 60 分 | 0.800 | 0.738 | 0.749 |
| 90 分 | 0.754 | 0.727 | 0.730 |
| 120 分 | 0.743 | 0.772 | 0.744 |
| 150 分 | 0.688 | 0.799 | 0.731 |

表 5　HDP-LDA の学習法の比較と単語事前分布の超パラメータ $\beta$ の適合率・再現率・$F$ 値に対する影響
Table 5　Comparison of learning methods and effects of hyperparameter $\beta$.

| 適用手法 | $\beta$ | 適合率 | 再現率 | $F$ 値 |
|---|---|---|---|---|
| CGS-HDP-LDA | $V/10$ | 0.527 | 0.764 | 0.615 |
| | $V$ | 0.800 | 0.738 | 0.749 |
| | $10V$ | 0.951 | 0.470 | 0.613 |
| CVB-HDP-LDA | $V/10$ | 0.441 | 0.730 | 0.537 |
| | $V$ | 0.830 | 0.629 | 0.682 |
| | $10V$ | 1.000 | 0.219 | 0.323 |

ならない．一部例外があるが，滑走窓の時間幅が長くなるほど，適合率が低下し，再現率が上昇する傾向がある．これは，滑走窓の時間幅が長くなるほど，1 回の通信に相当する基地局がより多くの Bag-of-Words ベクトルに出現することになり，検出されるクラスタ数が増加するためと考えられる．

表 5 は，HDP-LDA について，ベイズ学習法（CGS，CVB）と単語事前分布の超パラメータ（$\beta = V/10, V, 10V$）について比較した結果である．ベイズ学習法に関しては CGS がよいことが分かった．CVB は，初期値依存性が強いため，最適な解に収束していない可能性がある．単語事前分布の超パラメータに関しては，一般に，小さくするとトピック数が多く，大きくするとトピック数が少なくなる傾向があるが，本実験により $V$ がよいことが分かった．

#### 4.4.4　滞留度に基づく自宅・職場の検出

最後に，滞留地点の中で特に重要な自宅及び職場の検出能力の評価を行った．各手法ごとに得られたクラスタに対して滞留度を算出し，その順番に並べ換えた．表 6 に，各手法で自宅及び職場に対応づけられたクラスタの順位を示す．表中の「—」は，自宅と職場のクラスタが一つになってしまうなどの理由から，自宅または職場のクラスタが滞留地点と対応づけられない場合である．HDP-LDA は，過半数の被験者について，自宅，職場が上位 1 位，2 位に出現し，残りの被験者についても，自宅，職場が 3 位以内であり，他の手法に対する優位性が確認できた．被験者 No.3，No.4 は自宅と職場が近い距離にある被験者だが，HDP-LDA の場合は，出現時間帯の違いをもとにそれぞれを別の有意位置クラスタとして検出することができた．このことは，位置情報履歴の時間局所性を利用することの有効性を示している．

## 5. むすび

本論文では，携帯電話による通話・通信時に取得されるような，空間的粒度が粗く，時間間隔が一定ではない，疎な位置情報履歴から個人の有意位置を検出する手法について検討した．時間的に近接するデータ区間を文書，データ区間に含まれる通信基地局を単語とみなして，LDA と呼ばれるトピックモデルを適用する手法を提案した．実データに適用した結果，従来法である Leader Algorithm や iGMM よりも精度良く個

人の有意位置を検出することができることが分かった．LDA に対してデータからトピック数を自動的に調節するように拡張した HDP-LDA も，従来法よりも精度がよく，最適なトピック数を選択した場合の LDA に近い精度を達成した．HDP-LDA を適用する場合，時間窓に関しては 1〜2.5 時間程度の幅の時間窓がよく，単語事前分布に対する事前分布は一様分布に設定するとよいことが分かった．

本研究においては時間局所的な基地局の分布は独立同分布に従うものとみなして，LDA や HDP-LDA を適用した．今後の課題として，時間的な連続性を考慮したクラスタリング [23]〜[25] を適用することも検討したい．また，地理情報や利用者のライフスタイルを考慮したモデル化や，学習の高速化も検討したい．

**謝辞**　日ごろ御指導を頂く（株）KDDI 研究所中島康之所長に感謝申し上げます．

## 文　献

[1] 大橋正良（編），“特集 ライフログ，” 情報処理，vol.50, no.7, pp.589–640, 2009.
[2] D. Ashbrook and T. Starner, “Learning significant locations and predicting user movement with GPS,” Proc. 6th International Symposium on Wearable Computers, pp.275–286, 2002.
[3] D. Ashbrook and T. Starner, “Using GPS to learn significant locations and predict movement across multiple users,” J. Personal and Ubiquitous Computing, vol.7, no.5, pp.275–286, 2003.
[4] M. Ester, H.P. Kriegel, Jörg Sander, and X. Xu, “A density-based algorithm for discovering clusters in large spatial databases with noise,” Proc. 2nd International Conference on Knowledge Discovery and Data Mining, pp.226–231, 1996.
[5] P. Nurmi and S. Bhattacharya, “Identifying meaningful places: The non-parametric way,” Proc. 6th International Conference on Pervasive Computing, pp.111–127, 2008.
[6] S. Isaacman, R. Becker, R. Cáceres, S.G. Kobourov, M. Martonosi, J. Rowland, and A. Varshavsky, “Identifying important places in people’s lives from cellular network data,” Proc. 9th International Conference on Pervasive Computing, pp.133–151, 2011.
[7] J.H. Kang, W. Welbourne, B. Stewart, and G. Borriello, “Extracting places from traces of locations,” Mobile Computing and Communications Review, vol.9, no.3, pp.58–68, 2005.
[8] 遠山緑生，服部隆志，萩野達也，“携帯電話の測位機能を用いた有意位置の学習，” 情処学論，vol.46, no.12, pp.2915–2924, 2005.
[9] M.A. Bayir, M. Demirbas, and N. Eagle, “Mobility profiler: A framework for discovering mobility profiles of cell phone users,” Proc. International Conference on Pervasive and Mobile Computing, vol.6, no.4, pp.435–454, 2010.
[10] A.L. Barabási, “The origin of bursts and heavy tails in human dynamics,” Nature, vol.435, pp.207–211, 2005.
[11] D.M. Blei, A.Y. Ng, and M.I. Jordan, “Latent Dirichlet allocation,” J. Machine Learning Research, vol.3, pp.993–1022, 2003.
[12] J. Sivic, B.C. Russel, A.A. Efros, A. Zisserman, and W.T. Freeman, “Discovering object categories in image collections,” Proc. IEEE International Conference on Computer Vision 2005, 2005.
[13] C. Zhen and J. Xu, “Solely tag-based music genre classification,” Proc. International Conference on Web Information Systems and Mining, pp.20–24, 2010.
[14] H. Zhang, B. Qiu, C.L. Giles, H.C. Foley, and J. Yen, “An LDA-based community structure discovery approach for large-scale social networks,” Proc. IEEE International Conference on Intelligence and Security Informatics, pp.200–207, 2007.
[15] Y.W. Teh, M.I. Jordan, M.J. Beal, and D.M. Blei, “Hierarchical Dirichlet Processes,” J. American Statistical Association, vol.101, no.476, pp.1566–1581, 2006.
[16] T.L. Griffiths and M. Steyvers, “Finding scientific topics,” Proc. National Academy of Sciences, vol.101, pp.5228–5235, 2004.
[17] Y.W. Teh, D. Newman, and M. Welling, “A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation,” Advances in Neural Information Processing Systems, 2006.
[18] 上田修功，山田武士，“ノンパラメトリックベイズモデル，” 応用数理学会誌，vol.17, no.3, pp.196–214, 2007.
[19] L. Sethuraman, “A constructive definition of Dirichlet Priors,” Statistica Sinica, vol.4, pp.639–650, 1994.
[20] Y.W. Teh, K. Kurihara, and M, Welling, “Collapsed variational inference for HDP,” Advances in Neural Information Processing Systems, vol.20, 2008.
[21] 総務省編，平成 23 年版情報通信白書，2011.
[22] japan.internet.com 編集部，“1 日に送受信する携帯メールは平均何通？，” インターネットコム株式会社，http://japan.internet.com/research/20100331/1.html, 参照 Nov. 5, 2011.
[23] D. Blei and J. Lafferty, “Dynamic topic models,” Proc. 23rd International Conference on Machine Learning, pp.113–120, 2006.
[24] T. Iwata, T. Yamada, Y. Sakurai, and N. Ueda, “Online multiscale dynamic topic models,” Proc. 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.663–672, 2010.
[25] K. Ishiguro, T. Yamada, S. Araki, and T. Nakatani, “A probabilistic speaker clustering for DOA-based diarization,” Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, pp.241–

244, 2009.

（平成 23 年 7 月 1 日受付，11 月 7 日再受付）

**黒川　茂莉**　（正員）

2005 慶大・理工・管理卒．2007 同大大学院理工学研究科修士課程開放環境科学専攻了，同年 KDDI（株）へ入社．現在（株）KDDI 研究所 Web データコンピューティンググループに所属．ユーザモデリング，コンテクスト推定の研究に従事．2010 本会学術奨励賞受賞．情報処理学会会員．

**横山　浩之**　（正員）

1992 京大･工･電子卒，1994 同大大学院修士課程了．同年国際電信電話（株）（現，KDDI（株））へ入社．以来，研究所にて，ATM 網，移動通信網，IP 網，光パケット網の性能評価・設計に関する研究に従事．2004 より，クライアント及びサーバのプラットホーム構築に関する研究に従事．2000 本会学術奨励賞受賞．情報処理学会会員．博士（工学）．

**吉井　和佳**　（正員）

2008 京都大学大学院情報学研究科知能情報学専攻博士後期課程了．同年，産業技術総合研究所に入所し，現在に至る．音楽推薦システムや複数基本周波数推定など，機械学習に基づく音楽情報処理の研究に従事．山下記念研究賞，船井研究奨励賞など受賞．博士（情報学）．

**麻生　英樹**　（正員）

1981 東大・工・計数卒，1983 同大大学院工学系研究科修士課程情報工学専攻了，同年電子技術総合研究所入所．1993 年 9 月～1994 年 8 月ドイツ国立情報処理研究所客員研究員．現在，産業技術総合研究所知能システム研究部門主任研究員．統計的データ解析及び学習能力のある知的情報処理システムの研究に従事．