社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report
IBISML2010-65(2010-11)

ポスター講演

疎な位置情報履歴からの有意位置抽出方式に関する検討

黒川　茂莉†　　横山　浩之†　　吉井　和佳††　　麻生　英樹††

† 株式会社 KDDI 研究所 〒 356-8502 埼玉県ふじみ野市大原 2-1-15
†† 独立行政法人産業技術総合研究所 〒 305-8568 茨城県つくば市梅園 1-1-1 中央第 2
E-mail: †{mo-kurokawa,yokoyama}@kddilabs.jp, ††{k.yoshii,h.asoh}@aist.go.jp

あらまし　本稿では，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抽出する手法について検討した．時間的に近接する位置情報履歴を文書，各位置情報履歴に含まれる通信基地局 ID を単語とみなして，HDP-LDA と呼ばれるノンパラメトリックな文書ートピック確率モデルを用いて有意位置を抽出する手法を提案し，実データに適用した結果，精度よく個人の有意位置を抽出できることが分かった．

キーワード　位置情報履歴，クラスタリング，有意位置，トピックモデル，HDP-LDA

A Study of Extracting Personally Meaningful Places from Sparse Location Histories

Mori KUROKAWA†, Hiroyuki YOKOYAMA†, Kazuyoshi YOSHII††, and Hideki ASOH††

† KDDI R&D Laboratories Inc. 2-1-15 Ohara, Fujimino, Saitama 365-8502 Japan
†† National Institute of Advanced Industrial Science and Technology
AIST Tsukuba Central 2, 1-1-1 Umezono, Tsukuba, Ibaraki 305-8568 Japan
E-mail: †{mo-kurokawa,yokoyama}@kddilabs.jp, ††{k.yoshii,h.asoh}@aist.go.jp

Abstract　In this paper, we investigate methods to estimate personally meaningful places from geometrically sparse location histories which can be obtained from users' call or communication histories using mobile phones. We found that we can accurately estimate meaningful places by applying a topic model called HDP (Hierarchical Dirichlet Process) - LDA (Latent Dirichlet Allocation) to bags of cell IDs in time-windowed segments of histories.

Key words　location histories, clustering, personally meaningful places, topic models, HDP-LDA

1. はじめに

携帯端末で取得できる位置情報からその場所のその人にとっての意味を解釈する技術は，個人の状況に合わせた情報提供サービスを実現するための基盤的データを構成する上で重要である．ユーザ個人にとって特別な意味を持つ一定の空間領域は有意位置 [1], [2] と呼ばれる．本稿では、携帯端末で取得できる位置情報からユーザにとっての有意位置を抽出することを目的とする研究について述べる．

有意位置に基づくサービスを実現するうえで，位置情報の量・質とコストはトレードオフ関係にあり，データの量・質を高めようとすると，コストが増大する．例えば，GPS (Global Positioning System) による測位を頻繁に行うことで空間および時間の精度や粒度を高めることが可能になる反面，電力の消費量は増加する．特に，携帯電話の GPS 機能を利用する場合は，電池の消耗による待ち受け時間の減少に加えて，測位に起因するパケット通信の料金も利用者にとって負担になっていた．一方，有意位置の推定のために携帯電話での通信発生時に基地局側で取得可能な位置情報を用いる方法も考えられる．この場合，携帯電話端末側では何もする必要がないというメリットはあるものの，取得できる位置履歴情報の量・質が低下するという課題があった．

定期的な位置測位や GPS による高密度なデータを用いて位置情報をクラスタリングする方式としては，従来から，DBscan (Density-Based Algorithm for Discovering Clusters in Large Spatial Databases) を用いた例 [4] や GMM (Gaussian Mix-

ture Model) を用いた例 [5] があった．しかし，これらの方式は，通信時の基地局情報のような取得時間間隔が一定ではなく，空間的にも粒度が粗いデータに対して適用することが難しい．特に，位置情報が基地局の位置へ量子化された時系列に対しては，携帯電話の空間的な移動よりも，むしろ基地局の固定的な配置パターンによってクラスタリングの結果が左右される可能性が高い．

そこで，本研究では，基地局側で得られるような時間的，空間的に疎な位置履歴情報から，サービスの個人化にとって必要な有意位置を精度よく推定する方法を提案し，実際のデータで評価する．

以降，2. では提案手法を，3. では実験方法を，4. では評価と考察を，5. ではまとめと今後の課題を述べる．

2. 提 案 手 法

本稿では，HDP-LDA を用いて，疎な位置履歴情報から個人にとって意味のある位置を抽出する方法を提案する．以下では、HDP-LDA の元となった LDA について述べ、次に HDP-LDA について述べる。さらに、それらを有意位置の抽出に適用する方法について述べる。

2.1 LDA と HDP-LDA

LDA (Latent Dirichlet Allocation) はディリクレ分布と多項離散分布を活用した文書集合の階層ベイズ的な生成モデルであり，各文書中の単語の出現頻度のベクトル（Bag-of-Words ベクトル）に基づいて，文書データから文書中の潜在的なトピックを抽出する [3]．抽出されるトピックは，各単語が確率的に所属するクラスタであり，文書における各単語の出現の背後にある状況の一種と考えることができる．HDP (Hierarchical Dirichlet Process) -LDA は，階層ディリクレ過程を用いて LDA をノンパラメトリック化して，潜在クラス数がデータから自動的に決まるように拡張したモデルである [7], [8].

LDA では D 個の文書を含むコーパスデータの生成過程を以下のようにモデル化する（以下の定式化は [8] によるもので，オリジナルの [3] のものとは少し異なる）．

- トピックの単語分布 ϕ_z をディリクレ分布 Dir($\beta\tau$) からサンプル
- 以下を D 回繰り返して文書をサンプル
  - 文書の長さ（単語数）N_d をポアソン分布 Poisson(ξ) からサンプル
  - 文書のトピック分布 θ_d をディリクレ分布 Dir($\alpha\pi$) からサンプル
  - 以下を N_d 回繰り返して単語をサンプル
    - カレントトピック $z_{d,i}$ を多項離散分布 Mult(θ_d) からサンプル
    - 単語 $w_{d,i}$ をトピック $z_{d,i}$ ごとの単語分布（多項離散分布）Mult($\phi_{z_{d,i}}$) からサンプル

α, π, β, τ は超パラメータであり，それぞれ，文書ートピック分布の集中度と平均，トピックー単語分布の集中度と平均に対応する．Blei らは，これらの値を経験ベイズ法で最適化して一意に決定しているが [3] これらに対してさらに事前分布を設定することも行われている．事後分布や予測分布を解析的に求めることはできないため，変分ベイズ法による近似計算や，マルコフ連鎖モンテカルロ法による近似計算を行う．

LDA は文書データのモデルとして提案されたが，その応用範囲は広く，映画等のコンテンツのレーティングのモデル化や，コンピュータビジョンでの視覚情報処理などにも適用されている [6].

階層ディリクレ過程を用いて LDA における潜在トピック数を無限化し，事前にトピック数を決める必要を無くしたノンパラメトリックベイズモデルが HDP-LDA である．LDA と階層ディリクレ過程を組み合わせて拡張する定式化として，以下では [8] によるものを用いている．

HDP-LDA では，D 個の文書からなるコーパスデータの生成過程を以下のようにモデル化する：

- G_0 を ディリクレ過程 DP($\gamma, \beta\tau$) からサンプル
- 以下を D 回繰り返して文書をサンプル
  - 文書の長さ（単語数）N_d を Poisson(ξ) からサンプル
  - 文書のトピックの分布 G_d を DP(α, G_0) からサンプル
  - 以下を N_d 回繰り返して単語をサンプル
    - $G_d = \sum_{k=1}^{\infty} \theta_{d,k}\delta_{\phi_k}$ として，カレントトピック $z_{d,i}$ を無限次元の多項離散分布 Mult(θ_d) からサンプル
    - 単語 $w_{d,i}$ をトピック $z_{d,i}$ ごとの単語分布（多項離散分布）Mult($\phi_{z_{d,i}}$) からサンプル

超パラメータである α, β, γ については，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．τ は単語数次元の対称なディリクレ分布から生成する．

各文書ごとの G_d が G_0 をベースとする DP からサンプルされ，文書 d のトピック生起確率 θ_d が G_d によって決められる．従って，θ_d の次元は無限になりえる．各文書の G_d は共通の G_0 をベース測度としているため，その要素を共有している．このように，階層ディリクレ過程を使うことで，全体としてのトピック数を無限化しつつ，文書間でトピックが自動的に共有されるようなモデル化を行うことができる．

HDP-LDA も事後分布や予測分布を解析的に求めることはできないため，その計算には，変分ベイズ法やマルコフ連鎖モンテカルロ法が用いられる．実際の計算時には，十分に大きいトピック数の上限を定めて打ち切りを行う．HDP-LDA への変分ベイズ法の適用の詳細については [8] を参照されたい．

2.2 有意位置推定への適用

LDA や HDP-LDA を携帯端末と基地局との間の通信履歴データに適用するためには，文書および単語に相当するものを定義する必要がある．本研究では，時間幅 $= T$ ，シフト幅 $= S$ の滑走窓を用いて位置情報履歴から一定幅のデータ区間を切り出し，時間的に近接するデータ区間を一つの文書とみなす．さらに，各区間について，そこに含まれる基地局 ID，すなわち，その時間区間に通信を行った基地局 ID を単語とみなす．すなわち，各データ区間内に現れる基地局 ID の個数をカウントして Bag-of-Words ベクトルを生成して，LDA, HDP-LDA を適用した．このことは，各時間窓における基地局通信イベントが各人の潜在状態に依存するある確率分布に従って生成されたと

被験者 No.	1 日当たりのレコード数	基地局数
1	7.00	24
2	41.75	107
3	17.18	36
4	42.89	103
5	54.82	127
6	30.82	65
7	35.81	82
8	10.48	57
9	43.25	54
10	13.00	37
11	40.32	63
12	114.25	75

表 1　収集した基地局データ

仮定していることになる．

この手法の特長は以下の通りである．

- 時間間隔が一定でない位置情報履歴に対しても適用可能である
- 基地局の位置情報は用いていないため，空間的に疎な位置情報履歴に対しても適用可能である
- 同じ基地局に対しても，異なる時間帯であれば，異なるトピック＝有意状態が割り付けられる可能性がある

さらに，LDA や HDP-LDA を用いたクラスタリングの過程で計算されるパラメータを用いると，抽出されたトピックにおいてユーザが特定の位置に滞留しているか，それとも，特定の経路を移動中であるかの識別が可能である．滞留中には，少数（だいたい 3 地点くらい）の基地局とのみ通信を行う可能性が高く，移動中の場合には，多くの基地局と通信を行う可能性が高い．そこで，各クラスタ内での基地局分布のランダムさを表すエントロピーなどを計算し，その大きさをもとに滞留度を計算することができる．

3. 実 験 方 法

提案手法の有効性を検証するため，実際に携帯電話の通話・通信時に使われた基地局のデータに対して提案手法を適用して有意位置検出の精度等についての評価を行った．

3.1 データの概要

首都圏在住の計 12 名から，基地局の位置情報の履歴を 4 週間分取得した．並行して，検証用データとして，データ収集期間中の各被験者の滞留地点 (自宅，職場，および，各週のお出かけ先) についてのアンケートも行った．表 1 の通り，1 日当たりのレコード数（すなわち，通話やデータ通信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 未満であり，全期間中に観測された位置の数（基地局数）は 120 超，少ない人で 30 未満であった．

3.2 評価の方法

個人ごとのデータを，時間幅 $T = 60$ 分，シフト幅 $S = 15$ 分の時間窓で分割し，各データ区間内に出現した基地局をカウントして Bug-of-Words ベクトルを作成した．ベクトルの次元は，各個人ごとの 4 週間の位置情報履歴に現れた基地局の数である．この前処理されたデータに LDA, HDP-LDA をここで，各モデルの事前分布は無情報事前分布とし，事前パラメータは以下の通りとした．

- HDP-LDA のトピック集中度 α は 0.9 と 1 の間の乱数
- LDA の各単語のディリクレパラメータは 0.9 と 1 の間の乱数

その他のパラメータは [8] と同様に決めた．

比較のために，従来法である無限混合ガウス分布 (iGMM) を基地局の位置情報を含めた位置履歴情報に適用した．こちらは時間窓は切らずに，1 回の通信のあった基地局の緯度・経度を 1 つのデータ点として iGMM でモデル化している．iGMM の事前分布も無情報事前分布とした．

LDA, HDP-LDA の場合は，基地局のクラスタが得られ，各クラスタごとにそこに含まれる単語 (=基地局) の観測頻度の分布が得られる．LDA, HDP-LDA のクラスタリング結果に対する事後処理として，クラスタに対する基地局の割り当て数が全割り当て数の 1% を超えるかどうかという基準で，得られたクラスタの足切りを行った．すなわち，クラスタに対する基地局の割り当て数が全割り当て数の 1% 以下のクラスタは重要でないクラスタとみなし，抽出クラスタから除外した．

iGMM の場合は，基地局位置のクラスタが得られ，各クラスタごとに割り当てられる通信データの頻度が得られる．iGMM のクラスタリング結果に対する事後処理として，クラスタに対する通信データの割り当て数が全データ数の 1% を超えるかどうかという基準で，得られたクラスタの足切りを行った．すなわち，クラスタに対する通信データの割り当て数が全データ数の 1% 以下のクラスタは重要でないクラスタとみなし，抽出クラスタから除外した．

残ったクラスタについて，クラスタとアンケート結果の滞留地点の対応づけを行った．まず，各クラスタについて，滞留度 (LDA, HDP-LDA の場合はクラスタ内の基地局分布のエントロピーの逆数，iGMM の場合は第 1 主成分の分散の逆数) を計算した．次に，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履歴の緯度・経度に 2 次元正規分布を当てはめて信頼度=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 2km の円領域との結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対応づけを行った．棄却楕円の中心の周囲 2km の円領域と結合した理由は，棄却楕円が小さい場合は適切なクラスタであったとしてもアンケートで得た駅の緯度・経度との対応がとれない可能性があるためである．なお，正解滞留地点の優先度順は，被験者がアンケートの回答結果に記載した順そのままとした．正解滞留地点の優先度順をランダムに並べ替える等の操作により，対応づけに基づく評価結果が変化する可能性もある．

対応づけ結果に対して，次式で求められる適合率 (precision) と再現率 (recall) およびそれらから求められる F 値を計算して有意位置抽出結果の精度を評価した．

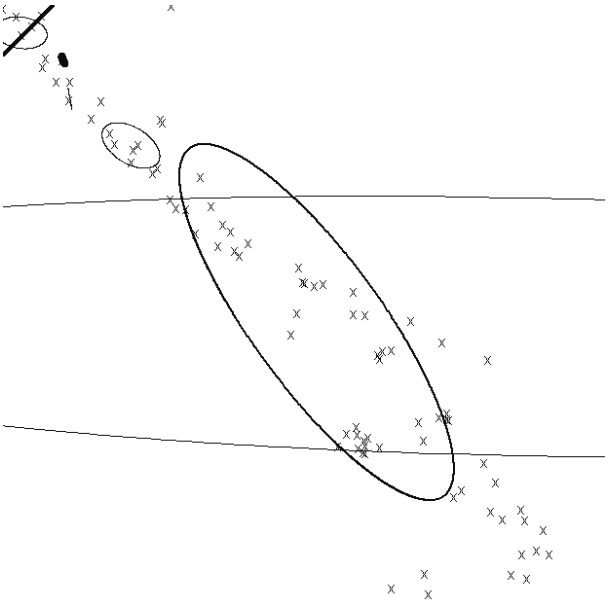

図 1　ある被験者の LDA によるクラスリング結果

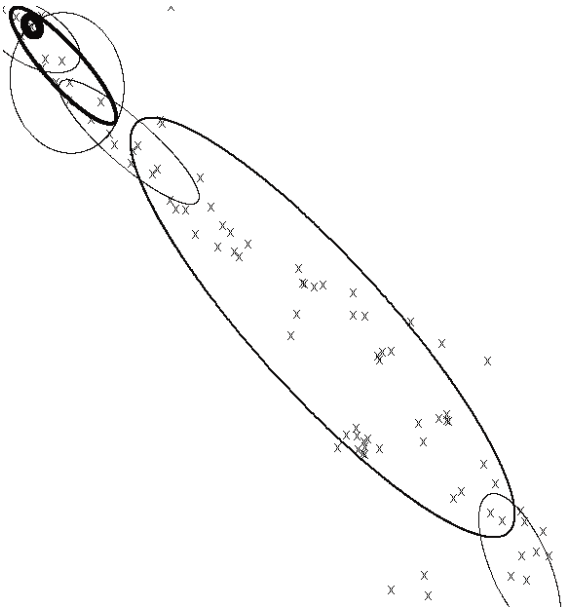

図 2　ある被験者の HDP-LDA によるクラスリング結果

$$\text{適合率} = \frac{\text{正解の滞留地点と対応がとれたクラスタ数}}{\text{足切り後のクラスタ数}}$$

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

また，滞留地点の中で特に重要な自宅，職場についての抽出能力の評価を行った．自宅，職場については，クラスタを滞留度の高い順に並び変えた場合，上位のクラスタと対応づけられることが望ましい．そこで，各手法のクラスタリング結果を，LDA, HDP-LDA の場合はエントロピー，iGMM の場合は第 1 主成分の分散で並び変えたときに，自宅，職場が何番目のクラスタと対応づけられたかを評価した．

4. 実験結果および考察

典型的な被験者について，LDA, HDP-LDA, iGMM で得られたクラスタの様子を図 1, 2, 3 に示す．ここでは，基地局または通信データの割り当て数が全割り当て数の 5% を超えるクラスタのみを表示している．図の中の黒い点は通信のあった基地局の緯度・経度を表し，楕円は抽出された各クラスタに対して前節で述べたようにして描いた 95% 棄却楕円であり，太い線のクラスタほど，滞留度が高いことを表している．これを見ると，LDA や HDP-LDA が大小様々なクラスタを生成しているのに対し，iGMM は少数の大きなクラスタを生成していることが分かる．図 1 では見難いが，図の左上方に，LDA の冗長なクラスタが密集している．

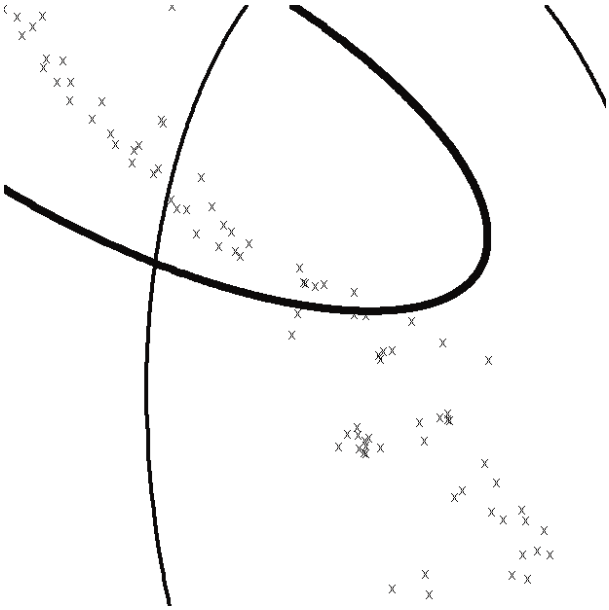

図 3　ある被験者の iGMM によるクラスリング結果

表 2 に，それぞれの手法による有意位置抽出の適合率，再現率，F 値の被験者ごとの値を示した．また，表 3 は自宅 (home), 職場 (office) に対応づけられたクラスタの順位を示している．

iGMM は領域の大きい少数のクラスタが生成されやすく，したがって，適合率は高いが，再現率が低い結果となった．クラスタ数が 1, 2 個の場合も多く，その場合は，職場に対応するクラスタがない被験者もあった．(表 3 の”-”で示している．) この原因としては，少数の点ではガウシアンパラメータの推定を安定させることが難しく，離れた位置を同じクラスタに含めてしまう傾向が強いことが考えられる．

それに対し，HDP-LDA は適切なクラスタサイズになっており，F 値において，多くの場合で iGMM よりよい結果を示した．観測される位置が離散的であるのに対して，素直に離散変数として扱っていることがよい結果に結びついたと考えられる．

LDA は，HDP-LDA と比べて，同様な情報をもつ冗長なクラスタが多数生成される傾向が強いため，適合率が低下している．また，自宅，職場のクラスタ順位においても他の手法と同等かそれ以下の結果となった．

5. おわりに

本稿では，携帯電話による通話・通信時に取得されるよう

	HDP-LDA			LDA			iGMM		
被験者 No.	precision	recall	F 値	precision	recall	F 値	precision	recall	F 値
1	0.40	0.40	0.40	0.24	0.80	0.36	1.00	0.20	0.33
2	0.67	1.00	0.80	0.32	0.75	0.44	1.00	0.25	0.40
3	0.78	0.78	0.78	0.21	0.44	0.29	1.00	0.22	0.36
4	0.69	0.82	0.75	0.53	0.91	0.67	1.00	0.27	0.43
5	0.58	0.58	0.58	0.42	0.67	0.52	1.00	0.17	0.29
6	0.60	0.75	0.67	0.26	0.63	0.37	1.00	0.38	0.55
7	0.70	0.78	0.74	0.47	1.00	0.64	1.00	0.33	0.50
8	0.75	0.25	0.38	0.20	0.25	0.22	1.00	0.17	0.29
9	0.83	0.83	0.83	0.21	0.67	0.32	1.00	0.50	0.67
10	1.00	0.56	0.71	0.44	0.78	0.56	1.00	0.22	0.36
11	0.56	0.71	0.63	0.26	0.71	0.38	1.00	0.14	0.25
12	0.45	0.71	0.56	0.26	0.71	0.38	1.00	0.57	0.73
平均	0.67	0.68	0.65	0.32	0.69	0.43	1.00	0.29	0.43

表 2　有意位置抽出結果の適合率，再現率，F 値

な，空間的粒度が粗く，時間間隔が一定ではない位置情報履歴から個人の有意位置を抽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 ID を単語とみなして，HDP-LDA と呼ばれるトピックモデルを適用する手法を提案し，実データに適用した結果，従来法のひとつである iGMM よりも精度よく個人の有意位置を抽出することができることがわかった．また，LDA による結果と HDP-LDA による結果を比較した結果，冗長なクラスタの抽出を抑制する点でモデルのノンパラメトリック化が有効であることがわかった．

今後の課題としては，提案手法のより詳細な評価およびモデルや手法の改良が挙げられる．まず，今回用いた有意位置抽出結果の評価に関しては，抽出されたクラスタと正解位置の対応づけの方法などに恣意性があり，改善の余地があると考えている．また，今回はユーザごとにクラスタリングを行ったが，有意位置を共有するユーザも多いと考えられることから，モデルを階層化して複数ユーザのデータをまとめて利用することが有効だと考えられる．また，ユーザのライフスタイルに関する情報など，位置情報履歴以外の情報も含めてモデル化してゆくことも興味深い．

	HDP-LDA		LDA		iGMM	
被験者 No.	home	office	home	office	home	office
1	2	1	4	1	1	-
2	1	2	1	2	1	2
3	1	2	1	2	1	-
4	1	3	1	2	1	3
5	3	1	2	1	1	-
6	2	1	8	1	2	1
7	2	1	1	2	2	1
8	1	3	-	1	1	2
9	2	1	1	2	2	1
10	1	5	1	3	1	2
11	2	1	7	1	1	-
12	2	3	1	10	1	2
平均	1.67	2.00	2.55	2.33	1.25	1.75

表 3　自宅 (home), 職場 (office) のクラスタ順位

文　献

[1] D. Ashbrook, T. Starner: Learning significant locations and predicting user movement with GPS, *Proceedings of the 6th International Symposium on Wearable Computers*, 275-286, 2002.

[2] D. Ashbrook, T. Starner: Using GPS to learn significant locations and predict movement across multiple users, *Journal of Personal and Ubiquitous Computing*, 7(5), 275-286, 2003.

[3] D. M. Blei, A. Y. Ng, M. I. Jordan: Latent Dirichlet allocation, *Journal of Machine Learning Research*, 3, 993-1022, 2003.

[4] M. Ester, H. P. Kriegel, Jörg Sander, X. Xu: A density-based algorithm for discovering clusters in large spatial databases with noise, *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, 226-231, 1996.

[5] P. Nurmi, S. Bhattacharya: Identifying meaningful places: The non-parametric way, *Proceedings of the 6th International Conference on Pervasive Computing*, 111-127, 2008.

[6] J. Sivic, B. C. Russel, A. A. Efros, A. Zisserman, W. T. Freeman: Discovering object categories in image collections, *Proceedings of the IEEE International Conference on Computer Vision 2005*, 2005.

[7] Y. W. Teh, M. I. Jordan, M. J. Beal, D. M. Blei: Hierarchical Dirichlet processes, *Techinical Report* December 15, 2005.

[8] Y. W. Teh, K. Kurihara, M, Welling: Collapsed variational inference for HDP, *Advances in Neural Information Processing Systems Vol.20*, 2008.