トピックモデルと大規模位置履歴を用いた地域ごとの興味関心分布の分析

宮永 薫[1,a)]　安納 爽響[1,b)]　坪内 孝太[2,c)]　下坂 正倫[1,d)]

概要：地域訪問者の興味関心分布の分析によって，訪問先の地域が持つ需要や特徴を明らかにすることができる．そのためこの分析は，訪問者の需要に合わせた店舗展開や，近年盛んに行われているユーザの訪問場所予測などに応用可能である．既存研究では，検索クエリと位置情報を用いた，訪問者の興味関心の分析が行われている．しかしこの研究では，訪問者の平均的な興味関心を捉えることはできるが，訪問者の多様性を表現できず，興味関心の異なる人が集まる大都市特有の傾向を明らかにできない．そこで本研究では，地域の訪問者を複数の訪問者タイプの集合として扱うことで，地域における興味関心の多様性を表現できる枠組みを提案する．具体的には，地域を文書，属性ごとの訪問者の興味関心ベクトルを単語，訪問者タイプをトピックと見立て，トピックモデルの一種である Gaussian Latent Dirichlet Allocation (GLDA) を活用することで，大都市における多様な興味関心の傾向や内訳を明らかにする．東京や大阪をはじめとする 9 都道府県における 6 ヶ月分の検索クエリと位置情報データを用いて，訪問者タイプとその内訳を抽出する性能評価実験を行い，提案手法が既存手法よりも大都市の分析に優れていることを示す．また様々な都市・POI ごとに興味関心分布を比較し，議論を行う．

1. 序論

地域の需要や特性を分析する研究は，従来より盛んに行われている．分析によって得た地域の需要や特性は，例えば効果的な店舗展開に関する商業活動 [1] や，地域ブランディングなどの都市開発 [2]，訪問者予測 [3] などに用いられている．このように地域の需要や特性の分析は幅広い分野への応用先が見込まれる研究と言える．

そういった地域需要や特性を分析する方法の一つとして，訪問者の興味関心を解析する手法が挙げられる [4]．特に，訪問者の位置情報を，彼らの普段の生活における検索ワードと紐付けた Sakamoto らの研究 [5] は，ユーザの普段の何気ない行動に垣間見える”興味”や”関心”と，地域の持つ特性とを紐づけて解析することができる研究として注目を集めている．

しかし，Sakamoto らの手法 [5] は，一つの地域における訪問者の興味関心に”多様性”を仮定しておらず，東京や大阪といった大都市のような，様々な興味関心を持つ人々が集まる傾向を捉えることができなかった．Sakamoto らの研究は，検索ワードから連想される 12 種類の興味関心軸をクラウドソーシングによって抽出し，ある地域と位置情報で紐づいた検索ワードを，TF-IDF を用いて 12 次元の興味関心ベクトルに変換している．その結果，地域の平均的な興味関心の傾向を掴むことに成功している．しかし，こういった興味関心の平均的な傾向は，都心や住宅地といった都市利用区分との相関を発見することを可能にしたが，都市の中に集まる訪問者の多様性を表現できず，興味関心分析の粒度に限界が存在する．

そこで本研究では，地域の訪問者を複数の訪問者タイプの集合として扱うことで，訪問者の多様性を表現することができる枠組みを提案する．具体的には，地域を文書，属性ごとの訪問者の興味関心ベクトルを単語，訪問者タイプをトピックと見立て，トピックモデル [6] の一種である Gaussian Latent Dirichlet Allocation（GLDA） [7], [8] を活用する．トピックモデルは文書解析のために発展した技術であり，文書解析に限らず幅広い分野へ応用されている．その中でも GLDA は，連続変数で表されたデータに適しており，我々が所望する枠組みに適している．

実験では，GLDA の優位性や，データの細分化による性能向上を平均対数尤度を用いて示す．またハイパーパラメータである訪問者タイプの数を perplexity を用いて調査する．さらに本研究では，東京や大阪をはじめとする 9 都

[1] 東京工業大学 情報理工学院 情報工学系
[2] LINE ヤフー株式会社
[a)] miyanaga@miubiq.cs.titech.ac.jp
[b)] anno@miubiq.cs.titech.ac.jp
[c)] ktsubouc@lycorp.co.jp
[d)] simosaka@miubiq.cs.titech.ac.jp

道府県における，6 ヶ月分の検索クエリと位置情報データを用いて考察を行う．実際の解析から得た結果をもとに東京や大阪の傾向を可視化し，地域の特徴や，抽出された訪問者タイプの特徴，POI の傾向に着目し，議論を行う．

本研究の貢献は以下のようにまとめられる．

- 本研究では，多様性に着目し，訪問者の日常的な興味関心を解析する枠組みを提案する．また枠組みを実現するために GLDA を用いた定式化を行う．
- 東京や大阪などの 9 都道府県における，6 ヶ月分の検索クエリと位置情報データを用いて，提案手法の訪問者の興味関心解析に対する性能について評価する．
- 解析結果を可視化することで，東京や大阪における地域の特徴や POI 分布の傾向を，実世界の特徴と合わせて議論する．

関連研究

訪問者数に着目した研究．携帯端末の発展などに伴い，訪問者数に着目して，地域の特徴を分析する研究が盛んに行われている．これらの研究では人流移動に着目し，特に都市動態に関する研究を行なっている [9]．また地域分析から応用し，イベント時の混雑予測にも発展されている [10]．これらの手法は，訪問者の増減が明確な地域 (オフィス街やレジャー施設など) の区別はつくが，一方で訪問者の増減があまりない地域の区別は難しい．

トピックモデルを用いた研究．地域の特徴分析のために，独自のトピックモデルを提案した研究が存在する [11], [12]．これらの研究では対象の地域を文書，地域にまつわるデータを単語，地域の利用法などをトピックとして分析を行なっている．一方で単語に当たるデータの形式に確率モデルの設計が大きく関わることから，汎用性は高いとは言えない．そのため，本研究が主眼とする訪問者の興味関心分析においては，活用が難しい．

地域を数値化する研究．近年盛んに行われている単語をベクトル表現で示す Word Embedding 技術を応用して，地域の情報を数値化する研究も行われており，訪問者予測や訪問地予測などの応用先を用いた評価で高い性能が報告されている [13], [14]．しかしながら，解析対象であるデータは訪問先の POI などに留まり，訪問者の興味関心を解析対象としているものは存在しない．

訪問者の興味関心に着目した研究．地域の特徴を捉えるため，訪問者の興味関心に基づいた研究が行われている [5], [15]．これらの研究は訪問者の興味関心を示すデータとして検索クエリを用い，都市を欲求という軸で表現した．さらに表現を用いた解析では，解析都市利用パターンを現実と近い精度で抽出した．一方でこの表現方法は，訪問者の多様性を捉えることができない．2 章にて詳述する．

2. 訪問者の興味関心解析の問題設定

本研究では地域訪問者の傾向を人の興味関心に基づいて解析する．具体的に本研究では地域訪問者の情報を数値化することが目的である．まず訪問者が訪れた地域を 1 辺 500m の正方形のメッシュに分割する．こうして D 個に分割されたメッシュを，**正方メッシュ地域**と呼ぶ．

2.1 GPS 情報と検索クエリによる訪問者の興味関心解析の既存手法

既存手法として，Sakamoto らが提案した，GPS 情報と検索クエリを用いた地域の数値化 [5] を挙げる．この枠組みでは，まず GPS 情報と検索クエリを訪問者を通して紐づけることで地域に基づいた検索クエリを抽出する．こうして得た検索クエリを構成する検索ワードの解析対象地域内での総検索回数，検索したユーザー数を調べ，検索ワードの重みを TF-IDF を用いて算出する．検索クエリをカテゴリに基づき 12 の欲求に表現したのち，検索ワードの重みを加味した地域内の検索クエリの平均を対象の地域における欲求スコアと題し，数値化する．12 の欲求とカテゴリの関係性は表 1 のようになる．

欲求	検索ワードのカテゴリ
同調欲	アイドル，ファッションモデル
好奇欲	テレビ，ニュース
生活安定欲	求人情報，職業，行政施設，公共施設
金銭欲	お金，通貨，投資，ギャンブル
物欲	製品名，店舗名，ファッション
服従欲	地域名，施設名
歓楽欲	文化，スポーツ，テーマパーク
知識欲	学校，教育
性欲	性的サービス，アダルト
生存欲	病気，怪我，出産，育児
怠惰欲	交通機関
食欲	料理レシピ，食材，飲食店

表 1: 検索ワードのカテゴリと欲求の対応

こうして得た欲求スコアを用い，土地利用パターンや利用者層を人口動態を用いた手法よりも高精度に推定し，地域の傾向との関連を分析した．

2.2 既存手法における訪問者の表現手法の限界

既存手法では 12 の欲求を軸とする数値を都市の特徴とみなしているが，算出過程をもとにするとその地域の平均的な訪問者像と見做せる．この場合，既存手法では地域の訪問者を一つの平均的な訪問者像のみを用いて表現していることとなる．しかし平均的な訪問者像のみでは訪問者の傾向を捉えられない地域が存在する．そう言った地域は，具体的には人の多く集まる都心などであり，なぜならば都心には，学生や社会人や主婦，またサブカルチャーに関心がある人，スポーツに関心がある人など，人の属性のみな

らず多様な興味関心を持つ人が訪れるからである．しかしSakamoto らの枠組みでは，こういった多様な訪問者の興味関心が平均化されてしまい，個性的な特徴も訪問者の多さゆえに薄れてしまい，得られる結果はどの関心やどの属性も平凡な訪問者像となる．以上のことから，既存研究は訪問者の多様性の表現に乏しい手法に留まっている．

3. 提案手法: 多様性に着目した訪問者の興味関心解析

本研究で提案する手法は，2.2 節で示した通り，既存手法で表現できなかった訪問者の多様性を表現するための枠組みと GLDA を用いた実現である．具体的には訪問者を複数の訪問者タイプに分類し，各地域ごとに訪問者タイプの内訳を抽出する．また属性ごとの興味関心ベクトルを単語，地域を文書，トピックを訪問者タイプとして扱うことで，トピックモデルの概念を取り入れ，GLDA を用いて本提案を実現する方法を示す．

3.1 トピックモデルを活用した訪問者の多様性の表現

トピックモデルは文書ごとにトピックが割り振られているという仮定に基づき，各文書のトピック，またトピックごとに登場頻度の高い語彙を抽出するための確率モデルの総称を指す．トピックモデルは一般に文書のラベルがついた単語を入力とし，各文書ごとのトピックの割合や各トピックごとの単語の出現割合を出力とする．

トピックモデルを活用するにあたり，本研究では新たに訪問者タイプという概念を導入する．これは訪問者が複数のタイプに分類できると仮定し，類似の訪問者を一つのタイプで統一的に扱うことができる．図 1 に文書分析と訪問者の対応を示す．本研究ではトピックモデルの入力として，地域ごとのラベルがついた興味関心ベクトルの情報を，出力として各地域の訪問者タイプの割合と各訪問者タイプごとの訪問者の分布とした．訪問者タイプはトピックモデルにおけるトピックにあたる．これによって各地域の訪問者の特徴だけでなく，複数の訪問者タイプを通して多様性を捉えることができる．

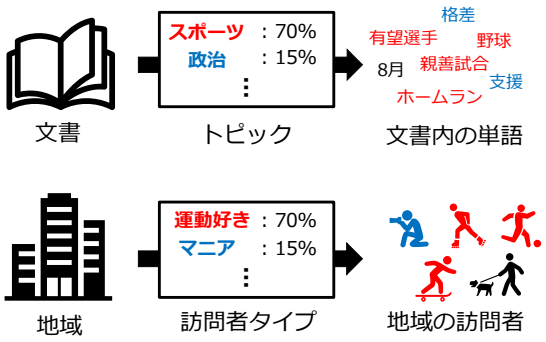

図 1: 文書解析との対応

3.2 Gaussian LDA による多様性を考慮した定式化

GLDA はトピックモデルの一種であり，単語が連続変数で表現されている文書の分析に適する [7], [8]．代表的なトピックモデルである LDA との違いは各トピックの単語分布がカテゴリカル分布ではなく，ガウス分布で表現している点である．これにより LDA では反映されなかった単語間の類似情報を用いた解析を行うことができる．本研究で扱うデータは 12 の数値化された情報で構成された連続データであることから，GLDA を選択する．

GLDA は確率分布に基づいた確率生成モデルである．そのため確率変数間の構造を図示するためのグラフィカルモデルを図 2 に示す．なおグラフィカルモデル内の枠は右下の数だけ内部の確率変数が存在すること，丸の内部が灰色である確率変数は観測変数であることを示す．

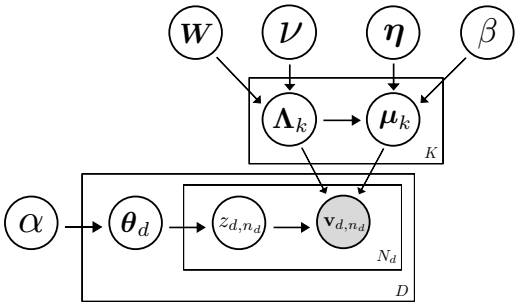

図 2: Gaussian LDA のグラフィカルモデル

生成過程に基づき，観測変数である興味関心ベクトル $\boldsymbol{V} = \{\boldsymbol{v}_{11} \cdots \boldsymbol{v}_{d,n_d} \cdots \boldsymbol{v}_{D,N_D}\}$ の生成までの手順を確認する．ここで興味関心ベクトルは M 次元のデータであり，地域の数を D，地域 d の興味関心ベクトルの数を N_d，訪問者タイプの数を K とする．まず初めに訪問者タイプの興味関心ベクトル分布を K 個生成する．訪問者タイプ k の興味関心ベクトル分布の精度 $\boldsymbol{\Lambda}_k$ は，パラメータ $\boldsymbol{W}_0$ と ν_0 を用いたウィシャート分布 $\mathcal{W}(\boldsymbol{W}_0, \nu_0)$ から生成する．生成された $\boldsymbol{\Lambda}_k$ と，パラメータ β_0 と $\boldsymbol{\eta}_0$ を用いた正規分布 $\mathcal{N}(\boldsymbol{\eta}_0, (\beta_0 \boldsymbol{\Lambda}_k)^{-1})$ から訪問者タイプ k の興味関心ベクトル分布の平均 $\boldsymbol{\mu}_k$ 生成する．ここまでで生成された訪問者タイプの精度と平均の集合をそれぞれ $\boldsymbol{\Delta}$，$\boldsymbol{M}$ とする．次に興味関心ベクトルを生成する．ディリクレ分布 $\mathrm{Dir}(\boldsymbol{\alpha}_0)$ から，地域ごとに訪問者タイプの割合 $\boldsymbol{\theta}_d$ を生成する．生成された $\boldsymbol{\theta}_d$ を用いて，生成する興味関心ベクトルがどの訪問者タイプに由来するかを決める潜在変数 $\boldsymbol{z}_{d,n_d}$ を $\mathrm{Cat}(\boldsymbol{\theta}_d)$ から生成する．最後に $\boldsymbol{z}_{d,n_d}$ に基づく訪問者タイプの興味関心ベクトル分布 $\mathcal{N}(\boldsymbol{\mu}_{z_{d,n_d}}, \boldsymbol{\Lambda}_{z_{d,n_d}})$ から興味関心ベクトル $\boldsymbol{v}_{d,n_d}$ が生成される．ここまでに生成された地域ごとの訪問者タイプの割合の集合を $\boldsymbol{\Theta}$，潜在変数の集合を $\boldsymbol{Z}$ とする．以上の生成過程で登場した各種変数の一覧を表 2 にまとめる．

生成過程に基づき，各変数の条件付き確率を示す．初めに全ての変数の同時布は $p(\boldsymbol{V}, \boldsymbol{Z}, \boldsymbol{\Theta}, \boldsymbol{M}, \boldsymbol{\Delta})$ は生成過程に従

名称	定義
データの次元	M
地域のインデックス	$d := 1, \cdots, D$
興味関心ベクトルのインデックス	$n_d := 1, \cdots, N_d$
訪問者タイプ	$k := 1, \cdots, K$
(d, n_d) の興味関心ベクトル	$\boldsymbol{v}_{d,n_d} \in \mathbb{R}^M$
興味関心ベクトルの集合	$\boldsymbol{V} = \{\boldsymbol{v}_{d,n_d}\}$
(d, n_d) に対する潜在変数	$\boldsymbol{z}_{d,n_d} \in \mathbb{R}^K$ where $\lvert\boldsymbol{z}_{d,n_d}\rvert_1, z_{d,n_d,k} \geq 0$
潜在変数の集合	$\boldsymbol{Z} = \{\boldsymbol{z}_{d,n_d}\}$
地域 d の訪問者タイプ分布	$\boldsymbol{\theta}_d \in \mathbb{R}^K$
地域ごとの訪問者タイプ分布集合	$\boldsymbol{\Theta} = \{\boldsymbol{\theta}_d\}$
訪問者タイプ k の興味関心ベクトル分布の平均	$\boldsymbol{\mu}_k \in \mathbb{R}^M$
訪問者タイプごとの興味関心ベクトル分布の平均の集合	$\boldsymbol{M} = \{\boldsymbol{\mu}_k\}$
訪問者タイプ k の興味関心ベクトル分布の精度	$\boldsymbol{\Lambda}_k \in \mathbb{R}^{M \times M}$
訪問者タイプごとの興味関心ベクトル分布の精度の集合	$\boldsymbol{\Delta} = \{\boldsymbol{\Lambda}_k\}$
地域 d に関するパラメータ	$\boldsymbol{\alpha}_0 \in \mathbb{R}^K, \alpha_{0,k} > 0$
訪問者タイプ k に関するパラメータ	$\beta_0 \in \mathbb{R}$ $\boldsymbol{\eta}_0 \in \mathbb{R}^M$ $\nu_0 > 0$ $\boldsymbol{W}_0 \in \mathbb{R}^{M \times M}$

表 2: 変数一覧

い $p(\boldsymbol{V} \mid \boldsymbol{Z}, \boldsymbol{M}, \boldsymbol{\Delta}) p(\boldsymbol{Z} \mid \boldsymbol{\Theta}) p(\boldsymbol{\Theta}) p(\boldsymbol{M}, \boldsymbol{\Delta})$ に分解できる．また分解された $p(\boldsymbol{V}|\boldsymbol{Z}, \boldsymbol{M}, \boldsymbol{\Delta})$，$p(\boldsymbol{Z}|\boldsymbol{\Theta})$，$p(\boldsymbol{\Theta})$，$p(\boldsymbol{M}, \boldsymbol{\Delta})$ を数式 1 から数式 4 に示す．

$$p(\boldsymbol{V} \mid \boldsymbol{Z}, \boldsymbol{M}, \boldsymbol{\Delta}) = \prod_d^D \prod_n^{N_d} \prod_k^K \mathcal{N}\left(\boldsymbol{v}_{d,n} \mid \boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1}\right)^{z_{d,n,k}}. \quad (1)$$

$$p(\boldsymbol{Z} \mid \boldsymbol{\Theta}) = \prod_d^D \prod_n^{N_d} \prod_k^K \boldsymbol{\theta}_{d,k}^{z_{d,n,k}}. \quad (2)$$

$$p(\boldsymbol{\Theta}) = \prod_d^D \mathrm{Dir}\left(\boldsymbol{\theta}_d \mid \boldsymbol{\alpha}_0\right). \quad (3)$$

$$\begin{aligned} p(\boldsymbol{M}, \boldsymbol{\Delta}) &= p(\boldsymbol{M} \mid \boldsymbol{\Delta}) p(\boldsymbol{\Delta}) \\ &= \prod_k^K \mathcal{N}\left(\boldsymbol{\mu}_k \mid \boldsymbol{\eta}_0, (\beta_0 \boldsymbol{\Lambda}_k)^{-1}\right) \mathcal{W}\left(\boldsymbol{\Lambda}_k \mid \boldsymbol{W}_0, \nu_0\right). \end{aligned} \quad (4)$$

4. 性能評価実験

本研究では，GLDA を活用し訪問者タイプの興味関心分布と正方メッシュ地域ごとの訪問者タイプの分布を生成するパラメータを推論することで，地域の訪問者を複数の訪問者タイプの重ね合わせで表す．そのため本実験は，複数の訪問者タイプの重ね合わせにより，訪問者をより高精度に表現できることを示すことを目的とする．

4.1 データセット

本実験で用いるデータセットは LINE ヤフー株式会社によって提供された city-atmosphere データセットである．このデータセットは位置情報や検索クエリから既存研究である Sakamoto らが提案した手法 [5] によって生成されたもので，各地域の特徴を月と訪問者の属性 (性別と年齢層) ごとに 12 の欲求の軸で数値化したものである．データセットは対象となる訪問者の属性，各欲求とその欲求スコア，トレンドワード，対象となる正方メッシュ地域の緯度経度が含まれているが，統計情報化されているためユーザーの特定が可能となる情報は含まれていない．解析対象地域は東京，大阪をはじめとする 9 都道府県であり，期間は 2022 年 3 月 1 日から 2022 年 8 月 31 日とした．本実験では GLDA の入力とするためにトレンドワードの数に着目し，データのフィルタリングを行う．最後に正方メッシュ地域ごとにデータを紐付け，その際に一定のデータ数を満たしたものを実験に用いた．

4.2 性能評価指標

モデルの推論により得られたパラメータを評価するため，平均対数尤度と perplexity の二つの指標を用いる．平均対数尤度 $p\left(\boldsymbol{V} \mid \mathcal{M}\right)$ は，GLDA の予測分布 $p(\boldsymbol{v}_{d,n}|\mathcal{M})$ を用いて以下のように表される．

$$p(\boldsymbol{V}|\mathcal{M}) = \frac{\sum_d^D \sum_n^{N_d} \ln p\left(\boldsymbol{v}_{d,n} \mid \mathcal{M}\right)}{\sum_d^D N_d}. \quad (5)$$

ここで，$\mathcal{M}$ はモデルを表し，$p(\boldsymbol{v}_{d,n}|\mathcal{M})$ は以下のように表される．

$$p(\boldsymbol{v}_{d,n}|\mathcal{M}) \simeq \frac{\sum_k^K \alpha_{d,k} \mathrm{St}\left(\boldsymbol{v}_{d,n}|\boldsymbol{\eta_k}, \boldsymbol{L}_k, \nu_k + 1 - M\right)}{\sum_k^K \alpha_{d,k}}, \quad (6)$$

$$\text{where } \boldsymbol{L}_k = \frac{\left(\nu_k + 1 - M\right)\beta_k}{1 + \beta_k} \boldsymbol{W}_k. \quad (7)$$

perplexity は言語モデルの評価に用いられる指標であり [6]，$\mathrm{perplexity}\left(\boldsymbol{V} \mid \mathcal{M}\right) = \exp\left(-p(\boldsymbol{V}|\mathcal{M})\right)$ と定義される． perplexity はモデルが隠された 1 単語の候補をどれだけ予測できるかを表し，値が小さいほど候補が少なく無駄なく予測できており，適したモデルであることを示す．

4.3 生成モデルにおけるベースライン (前提手法) の比較

本節では city-atmosphere データセットの分析における，GLDA の他モデルに対する優位性を示す．比較対象として Gaussian Mixture Model（GMM） [16] を用いる．GMM は一つのデータに対し，複数のガウス分布に属する確率を割り当てるソフトクラスタリングの一種である．本データセットの分析ではデータが属性ごとの興味関心ベクトル，各ガウス分布が訪問者タイプに対応する．GLDA との違いは正方メッシュ地域ごとの訪問者タイプ分布に違いがないことである．評価指標には平均対数尤度を用いる．

本実験ではトレンドワード数は 50 未満，1 正方メッシュ地域あたりのデータ数が 24 未満となるデータを省いた．1 正方メッシュ地域ごとに 4 つのデータをランダムに抜き出して作成したデータセットをテストデータ，残りのデータ

セットを学習データとし，データ数が約 1:8 の割合になるように設定した．抜き出しはランダムであるため，特定の属性のデータのみがテストデータ，または学習データに偏る可能性がある．

その後 GLDA， GMM ともに学習データを用いて学習を行い，学習データとテストデータとそれぞれの平均対数尤度を算出する．訪問者タイプの数は 10，終了条件は更新回数が 500 回になるまでと設定した．また学習は各モデルごとに 10 回行い，その平均を結果とした．

実験結果は表 3 に示す．GLDA は GMM よりも学習データでは約 0.49 ほど，テストデータでは約 0,36 ほど高く，GLDA の優位性を確認できる．

Model	**GLDA**	GMM
学習データ	7.45 ± 0.00	6.96 ± 0.00
テストデータ	**6.45 ± 0.00**	6.09 ± 0.00

表 3: 平均対数尤度によるモデルの性能比較.

4.4 平均値と細分化の比較

本節では訪問者分析において，属性別データが含まれたデータセットを学習に用いることの優位性を示す．比較対象は，GLDA を全属性の訪問者の平均を示すデータのみで学習した場合とし，これを Only Average と呼称する．Only Average の学習データは，city-atmosphere データセットから，すべての訪問者の平均を示すデータのみを抽出したものとした．テストデータには，4.3 節と同様，属性ごとの平均データが含まれたデータセットを用いる．評価指標は平均対数尤度を用い，実験設定は前節と同様に行った．

実験結果は表 4 に示す．属性別データが含まれたデータセットを学習に用いた場合の方が約 5.95 ほど高く，優位性を確認できる．

Training Data	**All Demographics**	Only Average
平均対数尤度	**7.36 ± 0.00**	1.41 ± 0.19

表 4: 学習データの違いによる性能比較.

4.5 トピック数の決定

本節では GLDA を用いて city-atmosphere データセットを分析する際に設定するハイパーパラメータの一つであるトピック数についての調査を目的とする．トピック数は訪問者タイプの数にあたり，ハイパーパラメータの中でも特に結果の有意性に大きく影響する．実験設定は 4.3 節で行った実験と同様の手法でデータの加工を行った後学習データとテストデータを作成する．比較手法として，perplexity を用いる．perplexity は 4.2 で述べた通り，隠された単語の候補の数を示す指標であり，一般に値が小さい方が性能が良いとされる．

実験結果を図 3 に示す．トピック数が増加することで perplexity が下がる傾向が見て取れる．これは訪問者タイプの数が増えることで，多くの訪問者に対応できることを示す．一方でトピック数が 20 を超えると，学習データに比べテストデータの perplexity の値に変化が確認できない．これは少数の訪問者を表現するための訪問者タイプが増え，トピック数が 20 以前と比べ予測精度への寄与率が低いためと考えられる．そのため最適なトピック数は 20 前後が適切であると結論づけられる．

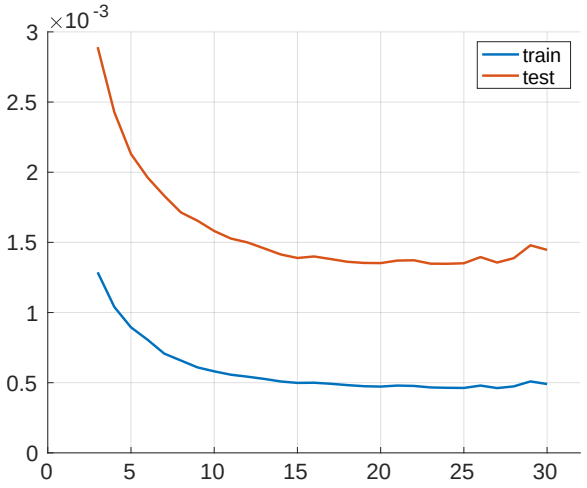

図 3: トピック数ごとの perplexity の比較

5. ケーススタディ：東京・大阪の解析結果の可視化

学習後に得たパラメータを用いて実世界でも特に東京や大阪の解析結果について議論を行う．議論のため，はじめに GLDA による学習によって得たパラメータを用いて正方メッシュ地域ごとの訪問者タイプ分布や，訪問者タイプの興味関心ベクトルの分布を算出する．まず正方メッシュ地域 d の訪問者タイプ分布 $\boldsymbol{\theta_d}$ は $\mathrm{Dir}(\boldsymbol{\alpha}_d)$ から生成した 10000 個の平均とした．次に訪問者タイプ k の欲求分布 k の分散 $\boldsymbol{\Lambda_k}$ と平均 $\boldsymbol{\mu}$ は $\mathcal{W}(\boldsymbol{W}_k, \nu_k)$ から分散を生成し，$\mathcal{N}(\boldsymbol{\eta}_k, (\beta_k \boldsymbol{\Lambda}_k)^{-1})$ から平均を生成する手順を 10000 回行い，得た値の平均とした．算出した正方メッシュ地域ごとの訪問者タイプ分布と訪問者タイプの興味関心ベクトルの分布の可視化は図 4 から図 15 と，表 5 のように示す．

5.1 東京・大阪における訪問者タイプの分布の可視化

抽出された訪問者タイプに関する可視化は図 4 から図 13 である．1 つの訪問者タイプにつき 2 種類のヒートマップと 1 つのレーダーチャートがある．ヒートマップは東京，大阪における正方メッシュ地域に訪れる訪問者タイプの割合を示す．黄色に近い正方メッシュ地域は対応する訪問者タイプに属する訪問者が多く，青色に近い正方メッシュ地域は対応する訪問者タイプが属する訪問者が少ないことを示している．レーダーチャートは訪問者タイプの興味関心ベクトル分布の平均を示す．色のついた線は対応する訪

問者タイプの値を，灰色の線はその他の訪問者タイプの値を示している．また表 5 では，抽出された訪問者タイプごとに，割合が高い上位 10 正方メッシュ地域で確認できた POI をまとめる．以下では抽出した訪問者タイプをヒートマップの特徴から，

- **一部の地域にのみ確認できる訪問者タイプ**：3，5，8
- **幅広い地域で確認できる訪問者タイプ**：1，4，7，9
- **郊外で確認できる訪問者タイプ**：2，6，10

の三つに分類し，議論する．

一部の地域にのみ確認できる訪問者タイプ

一部の地域にのみ確認できる訪問者タイプとして，訪問者タイプ 3，5，8 が挙げられる．これらの訪問者は特定の地域ではほとんど確認できない点も特徴である．

訪問者タイプ 3 は性欲と歓楽欲が高い訪問者タイプである．ヒートマップ，主要な POI の両方で東京ディズニーランドなどの総合娯楽施設が確認できる．性欲が高い理由として，デートなどの関連語が性欲として表れていると考えられる．訪問者タイプ 5 は怠惰欲が高い訪問者タイプであり，移動に関わる空港で割合が高くなっている．電車の乗換駅では確認できず，空港のみで確認できる理由として，電車よりも手軽さがない分，綿密に調べるための検索が増えるためと考えられる．訪問者タイプ 8 は歓楽欲や知識欲，服従欲が高く，都心で確認できる訪問者タイプである．一方で同調欲や好奇欲は低く，流行に敏感な若者ではなく，オフィス街のような地域に訪れる社会人のような訪問者を表していると考えられる．

幅広い地域で確認できる訪問者タイプ

幅広い地域で確認できる訪問者タイプとして，タイプ 1，4，7，9 が挙げられる．

訪問者タイプ 1 と 9 は各地域の割合が高い点，興味関心ベクトルが平均的である点などの共通点が確認できる．タイプ 1 はタイプ 9 と比べると生活安心欲 (+0.11) や金銭欲 (+0.13)，生存欲 (+0.11) が高い訪問者タイプであり，都心から離れた地域で割合が高くなる傾向があることが確認できる．このことからタイプ 1 の割合が高い地域では日々の暮らしにやや関心のある訪問者が多いと考えられる．一方でタイプ 9 は服従欲 (+0.1) が高く，都心から近い地域で割合が高くなる傾向がある．主要な POI に交通の便が優れた地域が多いことを踏まえると，住居と職場や学校が離れている住民が多いと考えられる．

訪問者タイプ 7 は抽出されたタイプの中で知識欲が最も高い訪問者タイプである．また路線沿いに割合が高くなる傾向がわずかに確認できる．主要な POI にはベッドタウンに関連する POI が多く，高度な仕事に従事する高所得者が多いと考えられる．一方で訪問者タイプ 4 は金銭欲と物欲が高い訪問者タイプである．ヒートマップでは確認が難しいが，主要な POI から競艇場や宝石加工の工場など，賭博や高価な商品に関連する POI がある地域で割合が高いことを確認できる．

郊外で確認できる訪問者タイプ

郊外で確認できる訪問者タイプとして，タイプ 2，6，10 が挙げられる．これらのタイプは全体に占める割合が高い地域でも 50%ほどだが，確認できる地域が一致していることから，郊外の訪問者をこの 3 種類の訪問者タイプでバランスよく表していると考えられる．

訪問者タイプ 2 と 6 は全ての欲求の中でも生存欲や食欲が高い傾向があるタイプである．歓楽欲や性欲といった欲求は低く，日々の生活に強い関心がある訪問者が多く訪れていると考えられる．その中でもタイプ 2 は特に生存欲と食欲が高く，主要な POI として病院が多く確認できた．これは病気になった際に検索する病院や，病気時の食事に関する事柄などが生存欲や食欲の高さとして反映されたと考えられる．タイプ 6 は性欲が極端に落ち込んでいる特徴も持つ．主要な POI として団地が多く確認でき，訪問者として性欲が落ちた年配層や子育てに従事する家族などが考えられる．

訪問者タイプ 10 は抽出された訪問者タイプの中でも同調欲や好奇欲が高い訪問者タイプである．流行に敏感であり，都心に関心のある訪問者を表していると考えられる．

5.2 POI ごとの訪問者タイプの内訳の比較

抽出された訪問者タイプを用い，POI ごとに比較しつつ訪問者タイプの内訳について図 14 から図 15 を用いて議論する．作成した帯グラフは対象の POI が含まれる正方メッシュ地域の訪問者タイプの割合を示す．対象の POI が複数の正方メッシュ地域にまたがっている場合は，その平均とした．また図で示す訪問者タイプは前節で議論を行った訪問者タイプと同一であることを留意していただきたい．

図 14 で様々な POI の分布を示す．銀座や渋谷駅，秋葉原駅は訪問者タイプ 8 の割合が高く，オフィス街に近い都心という特性が確認できる．その中でも秋葉原駅は金銭欲や物欲が高い訪問者タイプ 4 の割合が高い．これはサブカルチャーや電気工作の界隈において，秋葉原でしか入手できない商品の購入を目的とした訪問者が多いためと考えられる．一方で柴又や麻布，東京工業大学は訪問者タイプ 9 の割合が高く，住宅地という特性を確認できる．柴又は訪問者タイプ 2 や訪問者タイプ 7 など，日常の生活に興味のある訪問者が多いため，他の地域よりも住宅地という特性が強いと考えられる．麻布は訪問者 8 が高く，都心と住宅地の両方を持つ地域であると確認できる．最後に東京工業大学は訪問者タイプ 7 の割合が高い．大学としての特性があまり確認できていないのはデータ収集時期がコロナウイルスの流行と重なっており，学生の訪問が減っていた可能性がある．

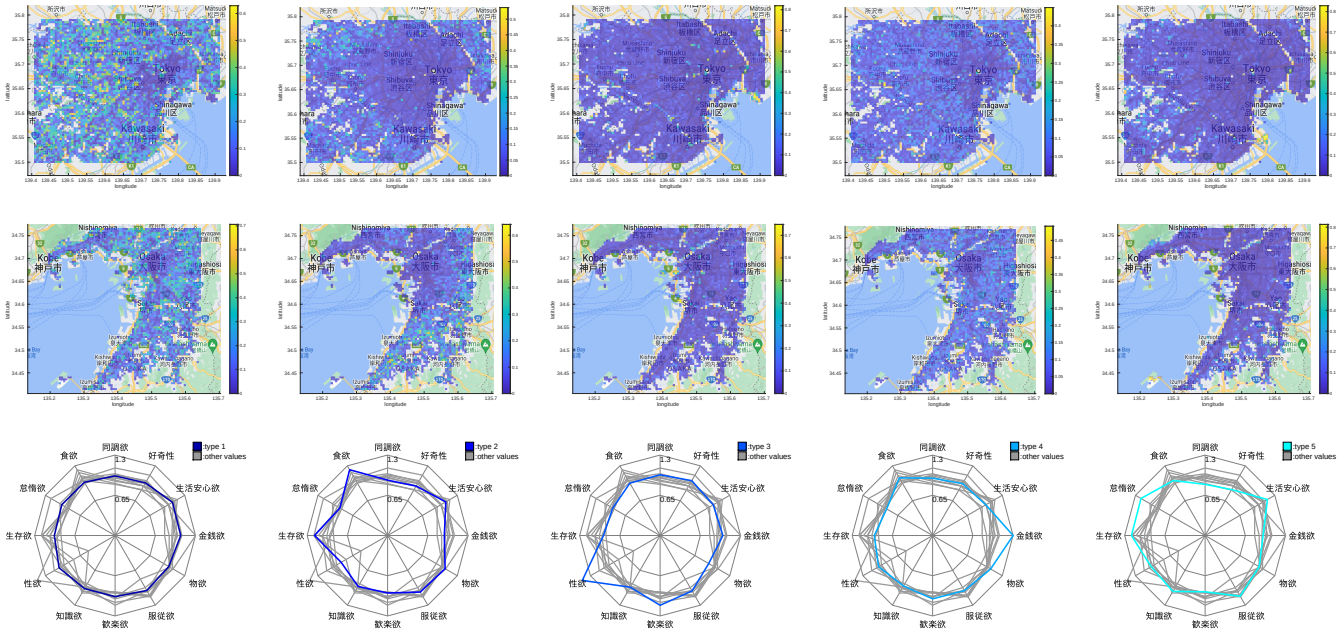

図 4: Type 1　図 5: Type 2　図 6: Type 3　図 7: Type 4　図 8: Type 5

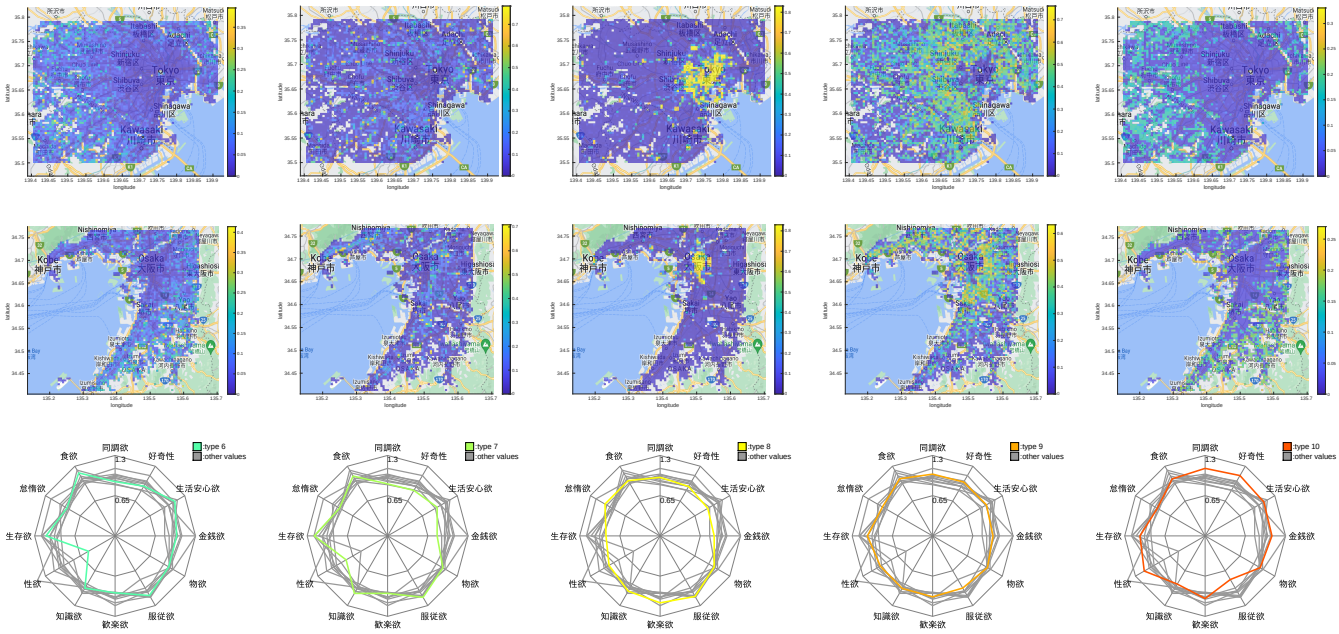

図 9: Type 6　図 10: Type 7　図 11: Type 8　図 12: Type 9　図 13: Type 10

また共通の特徴を持つ POI に注目して分析を行う．図 15 は主要な都心の駅の分布を示す．先ほど議論した秋葉原を除き，ほぼ同じ分布となっている．この結果は都心と駅という特性をあわせもつ POI は同様の分布であることが予測できる．図 16 は都内の大学の分布を示す．大学などの教育機関で確認できる訪問者タイプ 1 はどの大学でも 10%以上の割合である．特に東京大学や早稲田大学では 50%前後の割合を示している．大学の中でも分布の傾向が異なる理由は前述した通りコロナウイルスの流行や，キャンパスの立地による影響があると考えられる．

6. 結論

本研究では，既存の興味関心分析の枠組みにおいて，訪問者の多様性が表現されていないという課題に取り組み，Gaussian Latent Dirichlet Allocation(GLDA) を用いて地域訪問者を複数の訪問者タイプの割合で表現する手法を提案した．実験では GLDA が GMM よりも平均対数尤度において約 0.36 ほど高精度となること，訪問者を細分化することで全訪問者の平均のみを捉える場合よりも平均対数尤度で約 5.95 ほど高精度となること，最適なトピック数として 20 前後が最適であることを示した．さらに東京，大阪における解析結果を可視化し，POI と訪問者の興味関心の

抽出されたタイプ	主要な POI
タイプ 1	大阪学院大学, 甲南女子大学, 雲雀丘学園小学校, 大阪産業大学, 日本大学 関西学院大学, 学習院大学, 八街市立朝陽小学校, 東京都立武蔵野北高
タイプ 2	聖隷佐倉市民病院, 北里大学メディカルセンター, みつわ台総合病院, 千葉西総合病院 社会医療法人三栄ツガサキ病院, 千葉愛友会記念病院, 医療法人徳洲会羽生総合病院, 西狭山病院
タイプ 3	東京ディズニーランド, 東京ディズニーシー, J-GREEN 堺 町田 GION スタジアム, 聖丘カントリークラブ, ZOZO マリンスタジアム, ユニバーサルスタジオジャパン
タイプ 4	鷲宮ガス, ボートレース三国, 新日本工機信太山工場, エービーイーダイヤモンド, サンマル工業, ファイブカラット, ボートレース多摩川 笠置キャンプ場, かし原越谷工場
タイプ 5	羽田空港, 大阪国際空港, 成田空港, 関西国際空港, 神戸学院大学
タイプ 6	香日向団地, 西神ニュータウン、富岡団地, 鹿の子台団地, 葉山団地, 高松団地
タイプ 7	向ヶ丘遊園駅、寝屋川市駅, 洛北阪急スクエア, 中百舌鳥駅 浦安駅, 横須賀中央駅、仲町台駅、茨木市駅、三国ヶ丘駅
タイプ 8	京都市役所前駅, 東京丸の内, 北新地駅, 豊島ヶ丘市役所、上野駅 品川駅, 三宮駅, 恵比寿駅, 梅田駅周辺
タイプ 9	新宿 7 丁目周辺, 台東三筋周辺, 綾瀬駅, 三鷹上雀連周辺, 町屋駅周辺
タイプ 10	京浜島, 京葉市川 PA, 日本製紙草加工場, 日当物産、高谷 JCT, 北足立市場, 蓮田 IC, 美女木 IC

表 5: 各トピックの上位 10 正方メッシュ地域の主要な POI

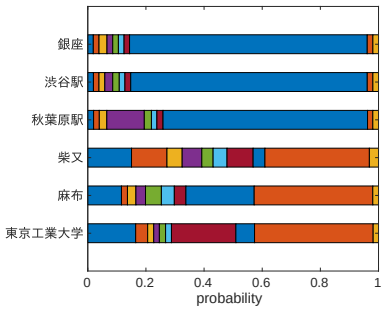

図 14: 様々な地域の分布

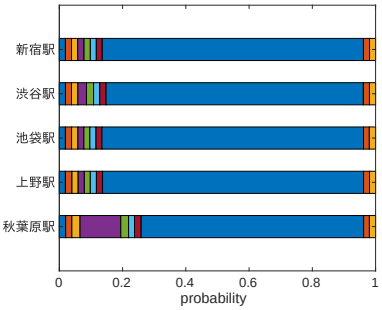

図 15: 様々な駅の分布

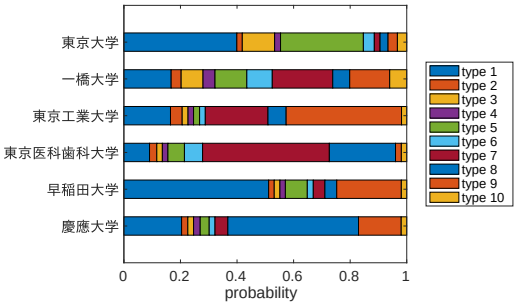

図 16: 様々な大学の分布

関係性を示した．将来課題として，地域のグルーピングや時系列分析を目的としたモデルの拡張や，訪問者一人一人に着目した分析，また欲求という表現にとらわれない興味関心の表現方法の構築が挙げられる．

参考文献

[1] Biswas, S. et al.: Analysis of different inventory control techniques: a case study in a retail shop, *Journal of Supply Chain Management Systems* (2017).
[2] Perkins, R. et al.: Understanding the contribution of stakeholder collaboration towards regional destination branding: A systematic narrative literature review, *Journal of Hospitality and Tourism Management* (2020).
[3] Ying, J. J.-C. et al.: Urban Point-of-Interest Recommendation by Mining User Check-in Behaviors, *In Proc. of SIGKDD* (2012).
[4] Rizwan, M. et al.: Visualization, Spatiotemporal Patterns, and Directional Analysis of Urban Activities Using Geolocation Data Extracted from LBSN, *ISPRS International Journal of Geo-Information* (2020).
[5] Sakamoto, T. et al.: CityAtmosphere: VR image to glimpse wishes in the air, *In Proc. of UbiComp* (2019).
[6] Blei, D. M. et al.: Latent dirichlet allocation, *Journal of machine Learning research* (2003).
[7] Hu, P. et al.: Latent topic model for audio retrieval, *Pattern Recognition* (2014).
[8] Das, R. et al.: Gaussian LDA for Topic Models with Word Embeddings, *In Proc. of ICNLP* (2015).
[9] Shimosaka, M. et al.: Spatiality Preservable Factored Poisson Regression for Large-Scale Fine-Grained GPS-Based Population Analysis, *In Proc. of AAAI* (2019).
[10] Hayakawa, Y. et al.: Simultaneous Multiple POI Population Pattern Analysis System with HDP Mixture Regression, *In Proc. of PAKDD* (2021).
[11] Yuan, J. et al.: Discovering regions of different functions in a city using human mobility and POIs, *In Proc. of SIGKDD* (2012).
[12] Kurashima, T. et al.: Geo topic model: joint modeling of user's activity area and interests for location recommendation, *In Proc. of WSDM* (2013).
[13] Feng, S. et al.: Poi2vec: Geographical latent representation for predicting future visitors, *In Proc. of AAAI* (2017).
[14] Yan, B., Janowicz, K., Mai, G. and Gao, S.: From itdl to place2vec: Reasoning about place type similarity and relatedness by learning embeddings from augmented spatial contexts, *In Proc. of SIGSPATIAL* (2017).
[15] Li, J. et al.: Capturing spatial distribution of people interests with web queries and location data: A large scale empirical study of metropolises in Japan, 研究報告ユビキタスコンピューティング (UBI) (2021).
[16] McLachlan, G. J. et al.: *Mixture models: Inference and applications to clustering*, M. Dekker New York (1988).