Vol.2019-DPS-179 No.9
Vol.2019-MBL-91 No.9
Vol.2019-ITS-77 No.9
2019/5/23

SNS上での言及を考慮した施設人気度の推定手法

川崎 仁嗣[1,2,a)]　榎園 健[1]　深澤 佑介[1]　豊田 正史[2]

概要：観光客への観光スポットの推奨や，施設名称での検索結果を提示する際に，施設（Point-of-Interest；POI）の人気度を考慮し，人気である施設を上位に提示することで，より多くのユーザーにとって期待に沿った提示内容となる．しかし，新たな施設の登場など人気度は時間経過によって変化するものであり，人による付与には多くのコストを要する．そこで，観光情報サイトやSNS上での施設に関する投稿情報を用いて施設人気度を推定することにより，これまでより低コストに施設人気度を付与でき，更新頻度を向上させることが可能となる．特にマイクロブログなどに代表されるSNSでの言及回数が多いPOIは話題になっていると考えられ，より大きい人気度を付与するべきである．このように、SNS上での言及数による人気度推定の方法が考えられるが，大量に存在するPOIのうちSNSで言及されるPOIは少なく，実際に調査したところ1割程度であった．残りの9割程度のPOIについても人気度を推定するため，SNS上で言及された場合の言及数をPOIのジャンルや所在地，周辺の滞在人口から回帰決定木を用いて推定する方法を提案する．SNS上で言及されている約2万件のPOIについて，実際の言及数を用いてモデルを学習し，その後，言及数が未知だったと仮定してモデルを用いて言及数を推定したところ，実際の言及数と高い相関があり，本提案手法により言及数を推定できる可能性が示された．

Point-of-Interest popularity estimation method considering references in social network services

Satoshi Kawasaki[1,2,a)]　Ken Enokizono[1]　Yusuke Fukazawa[1]　Masashi Toyoda[2]

1. はじめに

施設（Point-of-Interest；POIとも呼ばれる）とは，一般的にイメージされるような商業施設やオフィスビルなどの建物だけでなく，海岸や山などの自然物，観光名所，待ち合わせスポットやランドマークなど，ユーザーが関心を持つ場所を示すものであり，施設人気度とは施設に対する関心度の高さを表した度数である．多くのユーザーが関心を持つ施設には大きい人気度が，ごく一部のユーザーのみ関心を持つ施設には小さい人気度が付与されることを期待している．施設に対し人気度を付与されていれば，より多数のユーザーにとって関心のある施設情報をユーザーに提示することができる．例えば，観光客の滞在エリア周囲にある施設を推奨したり，地図サービスでの施設検索結果を提示したりする際に，人気度の高い施設を上位に提示することが考えられる．本研究では後者の例のように，施設検索を行う際の提示順を適切にするために施設に人気度を付与することを目的としており，ユーザーごとにパーソナライズした人気度を付与するのではなく，より多数のユーザーにとって人気である施設に高い人気度を付与する手法とした．施設検索において施設人気度を用いる主な理由としては，同名施設やジャンル名称での検索時の検索結果に順序付けをすることが挙げられる．具体例で述べると，検索キーワードとして「清水寺」を指定した場合，多くのユーザーは京都にある清水寺が検索結果上位に提示されることを期待する．しかし，「清水寺」は全国に数箇所存在するため，図1のように施設人気度が高い「清水寺」を上位に提示することが，多くの場合でユーザーが期待する順序となる．単純に距離が近い施設を提示する方法も考えられるが，「清水寺」などのように著名な観光スポットやランド

[1] 株式会社NTTドコモ
NTT DOCOMO, INC.
[2] 東京大学
The Uniersity of Tokyo
[a)] satoshi.kawasaki.vx@nttdocomo.com

マークの場合は人気度が高い施設を提示することが期待される．また，ジャンル名称での検索の場合は，「居酒屋」などの曖昧な検索キーワードの場合は検索結果となる施設が複数存在する．この場合も同様に，施設の人気度が高い居酒屋を上位に提示することが，多くの場合でユーザーが期待する順序であると考えられる．

施設人気度を付与する上で主に 2 つの課題が存在する．人気度は新たな施設の登場など時間経過とともに変化しており，人が施設の状況を調査して最新の人気度を随時反映するのは非常にコストがかかる．最新の状況を反映するために，より高頻度，例えば 1 週間に 1 回は人気度の更新を実施したいとすると，機械的に人気度を付与する仕組みが必要である．既存研究では，観光情報サイトにおけるコメント投稿や SNS 上での施設に対するチェックイン投稿などの情報から施設人気度を推定する方法 [1][2] が提案されている．一方で，このようなユーザーによる投稿情報を用いる場合，実際にチェックインの投稿が行われる POI は人気のある一部の POI に偏りがちである．このため，大多数の POI はほとんど言及されず，人気度の推定が出来ないという問題が挙げられる．

言及されない POI の割合を低減させるために，複数の観光情報サイトの情報を組み合わせたり，テキストだけでなく画像などの複数種類の情報を組み合わせたりする手法 [1] が提案されているが，それでもすべての POI が網羅されているわけではないため，人気度が付与できない POI が存在する．また，[2][3] では，SNS の投稿データのうちチェックインを示す投稿を用いているが，通常の投稿と比較しチェックインに関する投稿は件数が限られている．例えば，Twitter においては，位置情報付きの Tweet が全体の 0.58%[4] に過ぎず，また，Foursquare（現 Swarm）でのチェックイン情報を Twitter に連携しているユーザーは 15.7% [5] と少ない．チェックインに限定せず，POI に関して言及している投稿全体を利用することで，利用可能な投稿データ量が増えることから投稿データに紐づく POI の割合が増加することを期待できる．

本研究では SNS 上で言及されていない POI に対し，仮に言及されていたときにどれだけの言及数となるのかを回帰決定木モデルにより推定する手法を提案する．施設人気度の推定処理を図 1 に示す．提案手法は，「施設人気度生成」と「施設人気度推定」から構成される．施設人気度生成処理では，SNS の投稿データと POI データから施設人気度を生成する．次に，施設人気度推定処理では，生成した施設人気度を学習データとして推定モデルを構築し，SNS 上での言及がされていない POI の施設人気度を推定する．2 種類の施設人気度を組み合わせ、最終的な施設人気度とする．施設人気度生成処理は 3 章，施設人気度推定処理は 4 章で詳細を述べる．

SNS 上での言及がされていないことから，人気度を推定するために施設周辺の滞在人口や POI のジャンル，所在地の情報を特徴量として利用した．人口データは全国に整備された携帯電話の基地局から得られる運用データに基づき，あるエリアにおける端末在圏数から推計された人口 [6] を用いる．

本研究による貢献としては以下が挙げられる．

- SNS 上のチェックイン投稿に限らず，投稿データ本文に POI 名称や，その表記揺れ，人名・同一地名，チェーン名称を含んでいる場合に本来意図している POI への言及かどうかを判別することで施設人気度を付与した．
- SNS 上で言及されない POI に対しても，他の言及されている POI のデータを学習することで，施設周辺の滞在人口や POI のジャンル，所在地の情報から，仮に言及されたとしたときの言及数を推定する手法を提案した．
- 提案手法のモデルにより推定された言及数を，SNS 上での言及数と比較し，高い相関があることを確認した．

2. 関連研究

施設検索における検索結果提示順序の最適化を目的として施設人気度を算出している研究はあまり見られないが，一方で POI をレコメンドするために，ユーザーが関心のある POI を算出する研究は数多く見られる．

Yao ら [2] の研究では POI の人気度を POI 周辺 100m のエリア内における時間帯ごとの滞在人数と，POI のカテゴリーにおける時間帯ごとのチェックイン確率とを組み合わせる手法が提示されている．エリアの滞在人数についてはタクシーの降車位置と時刻のデータから集計し，カテゴリーごとのチェックイン確率については位置情報付き SNS の投稿データから集計しており，後者はカテゴリー単位での集計とすることでデータのスパース性を解決している．この手法では近隣エリアにある同一ジャンルの POI の場合，ほぼ同じ施設人気度が割り当てられるため，周辺にある居酒屋を人気度順でランキングにするような利用方法には適さない．本研究では，SNS 投稿データでの言及が存在しない POI については，POI のジャンルに加え，POI 周辺の施設数や性年代別の滞在人口などの特徴量から回帰決定木モデルによる推定を行う方法とした．

Ying ら [3] の研究では POI のジャンルごとにチェックインの頻度が異なることから同一ジャンルである POI のチェックイン数の合計に対する該当 POI のチェックイン数を施設人気度として用いることが述べられている．これにより訪問頻度が一般的に低いジャンルの POI（例えば海水浴場）であっても，他の訪問頻度が高いジャンルの POI と比較し大幅に低い施設人気度となってしまうことがなく，訪問頻度が高いジャンルの POI に偏ってレコメンドがなされてしまう問題が解決できる．本研究ではチェックイン

Vol.2019-DPS-179 No.9
Vol.2019-MBL-91 No.9
Vol.2019-ITS-77 No.9
2019/5/23

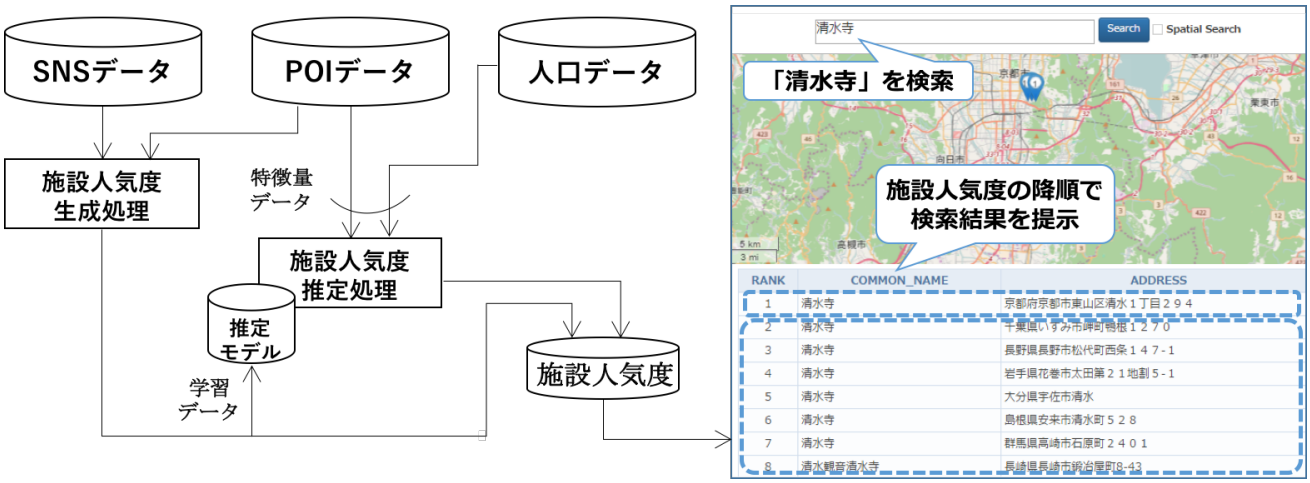

図 1 施設人気度の推定処理概要

ではなく，SNS での POI に関する言及ツイートの数を施設人気度として用いており，チェックインのしづらいジャンルであっても言及ツイートが少なくなるわけではないことから，ジャンル間での人気度の正規化は行わない．これにより，同一名称であるがジャンルが異なる POI であっても，より有名である POI に対して大きな人気度を割り当てることができる．

チェックインに関する投稿だけでなく，通常の投稿も用いて POI に対する言及がある投稿数を集計するには，投稿データの本文を解析して，POI の名称が含まれているかを確認する必要がある．POI の名称と同一の単語が含まれているからと言って，それが POI に対する言及とは限らないことがある．POI の名称が人名などの地名以外を指す言葉としても用いられていたり，地名だとしても同一の地名が複数箇所に存在したりする場合があり，実際に POI を指す言葉として用いられているかを判別する必要がある．このような投稿データ本文内の単語が地名を指すものとして扱われているのか否かを判別する手法として，Liu ら [7] の研究では条件付き確率場（Conditional Random Field；CRF とも呼ばれる）を用いた形態素の系列ラベリングを行うことが提案されており，落合ら [8] の研究では共起語を用いる手法が提案されている．

人気度を推定するにあたって，チェックイン数ではなく複数の観光情報サイトにおけるコメントや画像，レーティング情報を用いる手法 [1] も提案されている．本研究では人気度の推定対象となる POI の割合を増加させるために，投稿者数が限られる観光情報サイトではなく，より投稿者数が多い SNS の投稿を用いることとした．

3. SNS 投稿データからの施設人気度生成

SNS 投稿データ本文を用いた施設人気度は，それ自体を施設人気度として利用するだけでなく，言及がされない POI の言及数を推定するモデルの学習においても利用する．

施設人気度は，前章までに述べたとおり，チェックインに関する投稿だけでなく，通常の投稿も用いて POI に対する言及がある投稿数を集計することで算出を行う．実際にユーザーがその場所に行ったかどうかは考慮していないが，ユーザーが実際に訪問した，または，今後訪問しようと考えている POI であれば SNS へ投稿される可能性が高く，言及数は大きくなると考えられる．また，実際に訪問していなくても，施設に対する言及を行っているということは SNS 上で話題になっている，つまり，人気が高いことを反映していると考えられる．

投稿データの本文を解析して POI について言及しているかどうかを判断する上で，2 章で述べたように単純に POI の名称を含むかどうかだけでは判断できない．理由としては以下が挙げられる．

- 表記ゆれ問題
- 人名・同一地名問題
- チェーン総称問題

3.1 表記ゆれ問題

POI の正式名称の文字数が長い場合や，通称や愛称などが用いられることが多く，例として大学施設を考えた場合，大学の名称を略称で呼ぶことが多く，SNS の投稿データ本文においても略称で記載する可能性が高い．このような表記ゆれに対応するため，それぞれの POI についてあらかじめ略称などの別称を一覧化しておくことで，別称を含む投稿についても言及数として集計対象としている．

3.2 人名・同一地名問題

同じ POI の名称が複数ある場合や，POI の名称やその略称が人名など施設名以外の一般用語としても用いられている場合に，どの POI に対する言及であるかが曖昧となることである．例としては，「清水寺」は京都府京都市東山

区清水にある清水寺を指すこともあれば，千葉県いすみ市岬町鴨根にある天台宗の清水寺を指すことも考えられる．同様に，「宮島」は広島県廿日市市宮島町にある厳島の別称として知られるが，人名の姓としても利用される単語である．このため，投稿データ本文内に登場する単語が POI 名称やその別称と一致しているからといって単純に集計してしまうと正しい言及数にはならなくなってしまう．

落合ら [8] の研究では共起語を用いる方法が提案されており，従前の「清水寺」の例であれば，同じ本文内に「京都」,「舞台」などの単語が含まれる場合，京都府にある清水寺を指す可能性が高く，一方で「千葉」,「いすみ市」,「坂東三十三」などの単語が含まれる場合は千葉県の清水寺を指す可能性が高いと判断する．「宮島」の場合も同様に，「厳島」,「広島」,「神社」,「大鳥居」などの単語とともに用いられる場合は広島県の厳島を指す可能性が高く，「さん」などの単語があれば人名として用いられている可能性が高い．

本研究においては，POI に関する観光案内説明文を形態素分割し，単語出現頻度（TF と呼ぶ），逆文書出現頻度（IDF と呼ぶ）を求め，$TF \times IDF$ が一定のしきい値を超えるものを，該当の POI における共起語とした．投稿データ本文において，POI 名称やその別称とともに共起語を含むかどうかで言及の対象となっている POI を判別した．さらに，観光案内説明文が用意されていない POI については，CRF を用いて地名と推定された場合にのみ言及数として集計を行った．

3.3 チェーン総称問題

コンビニなど系列店舗が複数存在する POI の場合，特定店舗の POI に関する言及がほとんど行われなず，ユーザーはチェーン店であればどの店舗でも良い場合が多いため，具体的な店舗名まで含めた言及をすることは期待できない．本研究では，POI の名称からチェーン名称を抽出し，POI に対する言及ではなくチェーン名称に対する言及数を集計し，これを同一チェーンの各店舗の POI に均等に按分することで 1 店舗あたりの平均言及数を付与した．

4. 言及投稿がない POI の施設人気度推定方法

人気度のうち，3 章で述べた SNS データに基づく施設人気度が付与できる POI は，実際に SNS 上で言及されるような有名であったり話題になったりしている施設に限られるため，POI 全体のうち一部に限られる．およそ 3 ヶ月間の Twitter データに含まれる約 387 万件の Tweet を分析したところ，3 章の手法により POI を含むと判定された Tweet の件数は約 91 万件であり，言及の対象となっている POI の数は約 2 万施設であった．言及数の集計対象としている POI は約 18 万件であることから，Twitter データから言及数を集計できる POI はおよそ 1 割程度であると言える．図 2 は POI に対して言及している Tweet の数

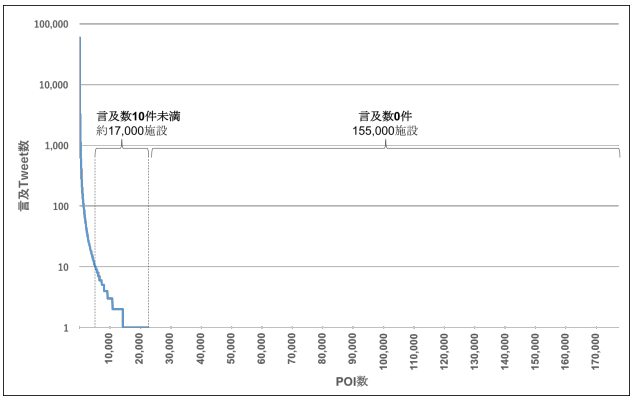

図 2 言及 Tweet 数のヒストグラム

を集計したものであり，10 件以上の言及 Tweet があるのは上位 3%の POI に過ぎず，それ以降は 10 件以下しか言及 Tweet がない POI がロングテールで存在し，さらに全く言及されない POI が 9 割ほどあることが分かる．

このように SNS データを用いる手法でのデータのスパース性については関連研究の多くで課題となっているが，本研究においては，言及されていない POI について，仮に言及されたときの言及数を回帰決定木のモデルによって推定する手法を提案する．

4.1 利用した特徴量

推定 POI 言及数の推定モデルでは，表 1 に示す 244 項目の特徴量を入力データとして用いる．

1) 施設ジャンル・業種

i 番目の POI P_i に対する施設ジャンル・業種 Cat_{P_i} は訪問者属性ごとの訪問頻度に影響すると考えられ，駅やコンビニのように訪問者属性間での差異が少ない施設ジャンルもあれば，学校や競馬場のように訪問者の年代に偏りが多い施設ジャンルもあるため，これらの違いを説明するための特徴量として用いる．

$$Cat_{P_i} = [G_1, G_2, \ldots, G_{10}, B_1, B_2, \ldots, B_6]$$

ここで，G_j は施設ジャンルであり 1 つの POI に対して最大で 10 個が存在し，B_k は業種を示すコードであり最大で 6 個が存在する．業種が 1 つであるがジャンルが複数ある POI も存在する一方で，業種とジャンルのいずれも存在しない POI もある．

2) 施設所在地

施設所在地 Loc_{P_i} は地域ごとの訪問傾向の差異を説明するための特徴量として用いており，POI が所在する地域メッシュ番号 $Mesh$，最寄り駅距離 $Dist_{ST}$，都道府県 JIS コード JIS_{PREF}，市区町村コード JIS_{MUNI} を含む．

$$Loc_{P_i} = [Mesh, Dist_{ST}, JIS_{PREF}, JIS_{MUNI}]$$

地域メッシュとは，総務省統計局が定めた一定の緯度経

Vol.2019-DPS-179 No.9
Vol.2019-MBL-91 No.9
Vol.2019-ITS-77 No.9
2019/5/23

表 1 推定モデルで用いた特徴量

特徴量種別	特徴量項目数	備考
施設ジャンル・業種	16 項目	一つの POI に複数ジャンルや業種が付与され得る
施設所在地	4 項目	地域メッシュ番号，最寄り駅距離，市区町村コードなど
周辺 200m 以内のジャンル別施設数	59 項目	例：小学校，宿泊施設（民宿）
属性ごとの平均滞在人口	117 項目	平休日，時間帯，性年代別
国勢調査における人口	46 項目	性年代属性別
平均滞在人口と国勢調査での人口との差分	1 項目	
POI データ取得元種別	1 項目	

度で区切られる矩形領域のことであり，本研究で利用したデータは 2 分の 1 地域メッシュで区切られている．具体的には 1 つのメッシュは $500m \times 500m$ の矩形領域となる．地域ごとの訪問傾向の差異とは，例えば，都心では駅の訪問頻度は高い一方で，郊外では自動車利用が多いことで駅の訪問頻度が低くなる地域もあると考えられる．最寄り駅距離は，電車を利用して施設に訪問するユーザーにとって訪問頻度に影響すると考えられ，同一の地域メッシュ内に同一ジャンルの施設があったとしても，駅からの距離がより近い施設が好まれる傾向が想定される．都道府県や市区町村コードは，同一の地域メッシュであっても行政区境界があることで，メッシュ内にある類似ジャンルの POI であっても訪問傾向が異なることを説明するために用いた．例えば，区役所の開庁時間が区ごとに異なっており，一方の区役所が閉庁していても別の区の区役所がまだ閉庁しておらず訪問者がいるような場合が想定される．

3) 周辺 200m 以内のジャンル別施設数

ジャンル別施設数 N_{P_i} は POI の周辺にあるジャンル別に集計した POI 数であり，本研究で利用したデータでは 59 個のジャンルが存在する．

$$N_{P_i} = [Q_{ct1}, Q_{ct2}, \ldots, Q_{ctl}]$$

ここで，Q_{ctl} は周辺 200m 以内でジャンル l である POI 数である．ジャンル別施設数は POI が存在する周囲の特性を説明するための特徴量であり，例えば宿泊施設や観光名所の POI が多いエリアは観光客が多いと考えられ，観光スポットなどの施設への訪問者が多いと考えられる．

4) 属性ごとの平均滞在人口

属性ごとの平均滞在人口 $Stay\_Pop_{P_i}$ は，[2] でのエリア活性度の考え方に類似しており，POI 周辺の平均滞在人口を集計した特徴量であり，平均滞在人口が多いエリアでは，より訪問者数が多くなると考えられる．また，POI のジャンルに応じて性年代ごとに訪問傾向が異なる可能性があるため，属性別に集計した平均滞在人口を用いる．これらの平均滞在人口は，前にも述べたとおり，携帯電話基地局の運用データに基づき，端末在圏数から推計された人口 [6] を用いており，ある時間帯の属性別人口をメッシュ単位で集計することができる．

$$Stay\_Pop_{P_i} = [S_{all}, S_{week}, S_{holiday}, S_{male}, S_{female}, S_{0,8,0,15}, S_{0,8,0,20}, \ldots, S_{d,t,g,a}]$$

ここで，S_{all} は全属性の合計平均滞在人口，S_{week} は平日の合計平均滞在人口，$S_{holiday}$ は休日の合計平均滞在人口，S_{male} は男性のみの合計平均滞在人口，S_{female} は女性のみの合計平均滞在人口である．また，$d \in \{0, 1\}$ は平日 (0) と休日 (1) を，$t \in \{8, 11, 15, 18\}$ は時間帯 (8 時〜20 時を 3 時間毎に 4 区分．ただし，11 時〜14 時のみ 4 時間) を，$g \in \{0, 1\}$ は男性 (0) と女性 (1) を，$a \in \{15, 20, 30, 40, 50, 60, 70\}$ は年代 (20〜70 才台まで 10 才区切りと 15〜19 才の 7 区分) を示している．

5) 国勢調査における人口

国勢調査における人口 $Census\_Pop_{P_i}$ は，国勢調査によって集計された人口の特徴量である．属性ごとの平均滞在人口と同様に，人口が多いエリアでは，より訪問者数が多くなると考えられる．国勢調査では居住地に基づき人口を集計しており，エリア内の居住人口を表しているが，属性ごとの平均滞在人口は昼時間帯におけるエリア内の人口であり，エリア外からの訪問者も含めた人口となっている点が差分である．また，国勢調査の場合，携帯電話を所持していない乳幼児や高齢者も集計されていることから，POI のジャンルによっては属性ごとの平均滞在人口よりも訪問傾向への寄与が大きくなることが期待される．

$$Census\_Pop_{P_i} = [C_{all}, C_{male}, C_{female}, C_{0,0}, C_{0,5}, \ldots, C_{g,a}]$$

ここで，C_{all} は全属性の合計人口，C_{male} は男性のみの合計人口，C_{female} は女性のみの合計人口である．また，$g \in \{0, 1\}$ は男性 (0) と女性 (1) を，$a \in \{0, 5, 10, \ldots, 90, 95, 100\}$ は年代 (0〜95 才台まで 5 才区切りと 100 才以上の 21 区分) を示している．

6) 平均滞在人口と国勢調査での人口との差分

平均滞在人口と国勢調査での人口との差分 $Diff_{P_i}$ は，属性ごとの平均滞在人口と国勢調査における人口との差分であり，エリア外からの訪問者のみの人口を表していると考えられ，宿泊施設のように居住者は訪問しづらいがエリ

表 2 評価におけるデータセット

データ種別	期間	件数
POI データ	2018.06 版	約 18 万施設
SNS(Twitter) データ	2018.6〜8	約 387 万件

ア外からの訪問者が立ち寄りやすい施設への訪問傾向を説明するための特徴量である．

$$Diff_{P_i} = [S_{all} - C_{all}]$$

7)POI データ取得元種別

POI データ取得元種別 $D\_Src_{P_i}$ は，POI データのデータソース種別であり，データソースごとの特性による人気度の差を説明するための特徴量である．本研究で用いている POI のデータソースは，観光者向け施設の POI データ，居住施設の POI データ，飲食店の POI データ，SNS 投稿データから抽出した POI データなどの種類が存在し，それぞれでジャンルは異なるものの同一名称の POI が含まれることがあり，かつ，ジャンルがどちらも付与されていないことがあった．実際にはジャンルが異なり人気度にも差があることから，人気度推定の際に識別可能とするための特徴量として用いた．

$$D\_Src_{P_i} = [Src]$$

ここで Src はデータソースごとにユニークに割り振られた識別子である．

4.2 推定モデル

推定モデルは，言及の投稿がある POI における言及数を正解値（目的変数）として，POI のジャンルや所在地など 4.1 節で述べた特徴量を説明変数に用いて学習を行う．

特徴量の次元数が高いこと，および，POI によりジャンルの付与数が異なるなど特徴量に欠損値があることから，このようなデータに対しても良好な精度を期待できるランダムフォレスト回帰決定木を用いた．

5. 評価

SNS での言及がなされていない POI に対して仮に言及されたとしたときの言及数の推定方法について 4 章で述べたが，本章では推定モデルの精度について評価を行った．評価に用いたデータセットの詳細を表 2 に示す．

推定モデルの学習および評価では，SNS データにおいて 1 回以上言及されている POI を用いて実施しており，対象となった POI は約 2 万施設である．言及数が既知の POI について，5 分割交差検証で推定精度の評価を行っており，ランダムフォレストの木の数は 20，500，1,000 個で実施して最も精度が高かった 1,000 個の場合で以降の評価を進めた．

評価指標として，言及数の真値との誤差を平均絶対誤差 (MAE)，平均二乗誤差 (RMSE)，相関係数で評価する．また，言及数により POI の提示順を決定する上で真値の言及数を用いた場合と推定モデルによる言及数を用いた場合とでの順位の誤差を nDCG(normalized Discounted Cumulated Gain) で評価した．推定モデルによる言及数を f_i，真値の言及数を y_i，評価対象とした POI 数を n としたとき，MAE は以下の式で定義される．

表 3 5 分割交差検証における精度

MAE	RMSE	相関係数	nDCG
7.41	10.52	0.60	0.90

$$\mathrm{MAE} = \frac{1}{n}\sum_{k=1}^{n}|f_i - y_i|$$

同様に，RMSE は以下の式で定義される．

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{k=1}^{n}(f_i - y_i)^2}$$

相関係数 cor は以下の式で定義される．1 に近づくほど，2 つの値の間の相関性が大きいことを示す．

$$cor = \frac{\frac{1}{n}\sum_{i=1}^{n}(f_i - \overline{f})(y_i - \overline{y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(f_i - \overline{f})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

ここで $\overline{f}$ は推定モデルによる言及数の平均値，$\overline{y}$ は真値の言及数の平均値である．

nDCG は以下の式で定義される．1 に近づくほど，2 つの順序関係の差異が少ないことを示す．

$$\mathrm{nDCG} = \frac{\sum_{r=1}^{n}\frac{2^{f(r)}-1}{\log_2(r+1)}}{\sum_{r=1}^{n}\frac{2^{y(r)}-1}{\log_2(r+1)}}$$

ここで r は順位，$f(r)$ は r 位の推定モデルによる言及数，$y(r)$ は r 位における真値の言及数である．

精度評価結果は表 3 に示すとおりであり，図 3 に POI 言及数（真値）と推定した POI 言及数との関係を散布図としてプロットした結果を示す．MAE が 7 前後，RMSE が 10 前後となっており言及数の推定誤差は無視できないが，相関係数で見ると 0.60 と比較的に相関が見られており，施設人気度での順序付けの差異については nDCG が 0.90 と高い値になっていることから，言及数に基づく施設人気度を用いて提示順の最適化を行う利用方法を想定すれば十分な精度であるといえる．以上より，POI データや人口データを用いた特徴量から POI 言及数を推定できる可能性が示された．

6. おわりに

本稿では，POI の人気度を推定するにあたって，SNS データを用いる手法では一部の有名な POI を除き大部分

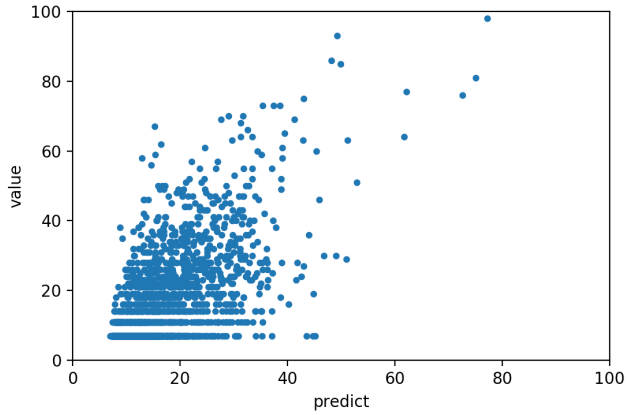

図 3 POI 言及数 (真値) と推定 POI 言及数の散布図

の POI が言及されないという問題に対し，チェックインに関する投稿だけでなく通常の投稿の本文を解析し POI への言及有無を判定することと，言及されていない POI が仮に言及されたときの言及数を回帰決定木のモデルによって推定する手法を提案した．

推定された言及数と，SNS 上での言及数とを比較評価した結果，相関係数が 0.60 と相関が見られ，施設人気度での POI 提示順序についても nDCG で 0.90 と真値での順序付けと近い結果が得られた．これまで言及されないことで人気度が推定できなかったおよそ 9 割の POI について本提案手法により言及数の推定値を付与することができた．

参考文献

[1] Y. Yang, Y. Duan, X. Wang, Z. Huang, N. Xie, and H. T. Shen. Hierarchical multi-clue modelling for poi popularity prediction with heterogeneous tourist information. *IEEE Transactions on Knowledge and Data Engineering*, Vol. 31, No. 4, pp. 757–768, April 2019.

[2] Z. Yao, Y. Fu, B. Liu, Y. Liu, and H. Xiong. Poi recommendation: A temporal matching between poi popularity and user regularity. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*, pp. 549–558, Dec 2016.

[3] Josh Jia-Ching Ying, Eric Hsueh-Chan Lu, Wen-Ning Kuo, and Vincent S. Tseng. Urban point-of-interest recommendation by mining user check-in behaviors. In *Proceedings of the ACM SIGKDD International Workshop on Urban Computing*, UrbComp '12, pp. 63–70, New York, NY, USA, 2012. ACM.

[4] Kisung Lee, Raghu K. Ganti, Mudhakar Srivatsa, and Ling Liu. When twitter meets foursquare: Tweet location prediction using foursquare. In *Proceedings of the 11th International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services*, MOBIQUITOUS '14, pp. 198–207, ICST, Brussels, Belgium, Belgium, 2014. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering).

[5] Yang Chen, Chenfan Zhuang, Qiang Cao, and Pan Hui. Understanding cross-site linking in online social networks. In *Proceedings of the 8th Workshop on Social Network Mining and Analysis*, SNAKDD'14, pp. 6:1–6:9, New York, NY, USA, 2014. ACM.

[6] 寺田雅之, 永田智大, 小林基成. モバイル空間統計における人口推計技術. NTT DOCOMO テクニカル・ジャーナル Vol.20 No.3. 一般社団法人電気通信協会, Oct. 2012.

[7] Xiaohua Liu, Shaodian Zhang, Furu Wei, and Ming Zhou. Recognizing named entities in tweets. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pp. 359–367, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[8] 落合桂一, 鳥居大祐. 時間変化する特徴語によるマイクロブログ地名曖昧性解消. 情報処理学会論文誌データベース（TOD）, Vol. 7, No. 2, pp. 51–60, Jun. 2014.