市民のツイートを利用した分散表現に基づく都市別特徴の可視化

安藤　有生†　　関　　洋平††

† 筑波大学 情報学群 知識情報・図書館学類　〒 305–8550 茨城県つくば市春日 1–2
†† 筑波大学 図書館情報メディア系　〒 305–8550 茨城県つくば市春日 1–2
E-mail: †s1411490@u.tsukuba.ac.jp, ††yohei@slis.tsukuba.ac.jp

あらまし　市民の移動は，都市ごとに異なった傾向にあり，年月が経過するとともに変化していく．そのような特徴をもつ市民の移動を把握することは，市民の需要を把握するうえで重要である．本研究では，リアルタイム性を持つ Twitter のデータを使用し，都市ごとに市民の移動の傾向を分析する手法を提案する．具体的には，ユーザのプロフィールに記述されている情報を利用し，都市ごとに収集した Twitter ユーザを用いて，地名と移動に関する動詞が出現しているツイートを，市民の移動先を表すツイートとして収集する．次に，市民の行動や認識の違いにより行き先となる地名を含むツイートの表現が異なると考え，市民の移動先を表すツイートをもとに，skip-gram を用いて単語の分散表現を生成し，多くの人がツイートしている行き先の地名を可視化する．可視化は，地名に対する市民の行動や認識を表す 2 つの単語と地名との単語の分散表現の類似度をそれぞれ Z スコアで正規化し，その差分を利用して行う．さらに，クラスタリングにより，市民の移動の傾向が似ている都市を分類する．提案手法の有効性について検証するために，8 つの都市ごとに収集した Twitter ユーザについて，市民の移動の傾向を分析する実験を行った．実験の結果，人気のある観光地，居住している都市，帰省先の地名が適切に可視化されており，観光目的や距離的に近い地名が同一クラスタにまとめられる傾向があることを確認した．

キーワード　Twitter，分散表現，市民の移動

1. はじめに

1.1 本研究の背景

市民の移動を把握することは，市民の需要を知るうえで重要なことである [18] [19] [27]．例えば，自治体は，市民の生活に合致した政策を実行することが求められている．そのためには，市民の生活の実態を把握する必要がある．市民の生活の一つとして，通勤・通学，余暇の移動があげられる．市民の移動に関する情報を収集する手段として，総務省が 5 年ごとにおこなっている国勢調査や，観光庁が 3ヶ月ごとにおこなっている宿泊旅行統計調査を使用することが考えられる [10] [22] [24]．しかし，国勢調査では最低でも 5 年前の実態しか知ることができず，現在も同じ状況が続いているかはわからない．また，都市ごとに市民の移動は異なる傾向がみられることから，都市ごとのデータが必要となる．国勢調査では，従業地，通学地に関する情報が，市区町村ごとに公開されている．一方で，宿泊旅行統計調査では，宿泊者数が都道府県単位で公開されている．このことから，自治体にとって現在の市民の移動の状態やその目的を把握することは難しいといえる．

本研究では，国勢調査や，人手と時間を必要とするアンケート調査やクラウドソーシングと比べて優れた，市民の移動を分析する手法を提案する．具体的には，リアルタイム性を持つ (注1) のデータを使用し，都市ごとに市民の移動を分析する．また，「行く」と「帰る」のような移動の向きや，「（過去に）行った」と「（将来）行きたい」のような認識の違いを分析できるよう，単語の分散表現 [20] を用いた手法を提案する．著者が調べた範囲では，都市ごとに Twitter ユーザを収集し，単語の分散表現の違いを用いて分析を行った研究は存在しない．

(注1)：https://twitter.com/

1.2 移動の分析

Twitter ユーザは，投稿の手軽さがゆえに，現在の状況や考えていることを投稿することが多い．観光庁 [31] [32] は，Twitter を用いて訪日外国人観光客に人気のある観光スポットを分析した．松本ら [25] は，Twitter を用いて，地域・観光情報の収集を行った．この研究で松本らは，Twitter について「特に速報性が高く，手軽に情報発信が可能」であると述べている．本研究では，市民の移動に関する情報が含まれているツイートを利用して分析を行う手法を提案する．

Twitter を用いて移動を分析する手法として，ツイート中に頻出する地名を抽出する手法が挙げられる [31] [32]．しかし，この手法だけでは，Twitter ユーザが興味を持っているために多く出現した地名は抽出できるが，抽出できた地名に実際に行ったのか，行きたいと考えているだけなのかといった分析はできない．頻出する地名を抽出する手法の他に，前後の文脈を考慮する，単語の分散表現を使用して抽出する手法が考えられる [14] [15] [26] [28]．

単語の分散表現とは，深層学習 [16] などを利用して，ベクトル空間内に単語を埋め込むことにより，単語をベクトルで表したものである．単語そのものを 1 つの成分とする，疎なベクトルで単語を表現した局所表現に対し，前後に出現する単語を考慮した単語の分散表現は，共起する単語に基づいて生成される．

これは，単語の意味は，共起する周辺の単語によって決まるという分布仮説 [3] [4] に基づいており，坪井ら [20] は，「単語の分散表現を用いることで，間接的に離散オブジェクト間の類似度や関係性を表現することができる」と述べている．また，中谷ら [13] は，単語の分散表現は，単語の意味を実数ベクトルで表したものであり，単語の近さはすべてベクトルの近さとして計算できるとしている．これにより，似た文脈で用いられる単語は，似た意味の単語と仮定でき，コサイン類似度が高く，近いベクトルで表されると考えられる．

単語の分散表現を作成する有名なツールに Tomas Mikolov ら [8] [9] によって提案された word2vec(注2) と，fastText [1] がある．word2vec と fastText の違いの一つに，fastText は subword を考慮する点があげられる．subword を考慮することにより，未知語の分散表現が高精度で獲得できるようになった一方，文字列が似ている低頻度語の分散表現が近くなる特徴がある．本研究では，低頻度の地名が多く出現する移動に関するツイートを対象にしているため，word2vec を使用する．word2vec は，skip-gram [9] と CBoW [8] の総称で，高速に単語の分散表現を作成できる．坪井ら [13] は，「大規模データによる単語の分散表現の学習が現実的な計算速度とメモリ量で実現可能になった」と述べている．skip-gram は，ボレガラら [21] は「対象語を使って文脈中に出現している文脈語を予測する」モデルであると述べており，ある単語の周囲に存在する単語の生起確率を推定するモデルである．都市ごとに，市民の移動を分析する際，出現した地名が，「行きたいと思ってはいるが行っていない場所」，「電車を用いて通勤している場所」，「旅行で行った場所」のどれに該当するかといった市民の認識を判別する必要がある．本研究では，市民の認識の違いにより，地域名を含むツイートの表現が異なると仮定し，市民が持つ，地名に対する認識の違いを分析するため，前後の文脈を考慮して生成される単語の分散表現を用いる．

開地ら [14] は，意味的に関連の強い観光地は，コサイン類似度の高い単語の分散表現を獲得できる点を可視化により示している．本研究では，分析の際，単語の分散表現により得られる 2 つの指標を考慮して分析を行う．1 つ目は，「行く」と「帰る」といった，移動の向きである．例えば，つくば市民の場合，秋葉原へは「行く」と言うのに対し，秋葉原にいる場合はつくばに「帰る」と言う．2 つ目は，「行った」と「行きたい」などの，市民の認識である．例えば，つくば市民の場合，秋葉原への距離は近いため，秋葉原へ「行った」市民が多いのに対し，距離が遠く頻繁に行くことが難しいが，魅力的な観光地である札幌へは「行きたい」市民が多いと考える．2 つの指標を複合的に分析するために，本研究では，2 つの指標をそれぞれの軸とする 2 次元に可視化を行い，視覚的に都市の傾向の類似性を把握できる手法を提案する．

都市別に，Twitter ユーザの移動について分析する手法の全体図を図 1 に示す．手法の詳細については，3 章で述べる．

本稿の構成を以下に示す．2 章では，SNS を使用した観光や

(注2)：https://code.google.com/p/word2vec/

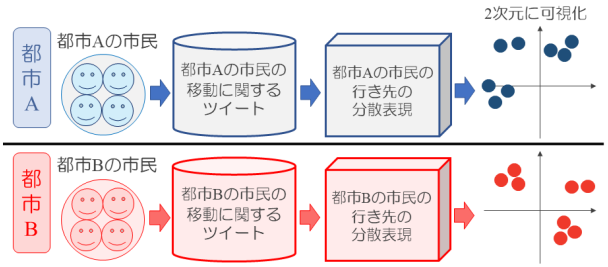

図 1　都市を単位とした市民の移動の分析

移動についての研究と，単語の分散表現を活用した研究について論じ，本研究の位置づけを明らかにする．3 章では，都市ごとに市民の移動を分析する手法を提案する．4 章では，3 章の提案手法の有効性を検証するために，実験を行い，5 章で考察する．6 章では，本研究で得られた知見をまとめ，現在の課題や今後の予定について述べる．

2. 関 連 研 究

本研究では，Twitter の投稿であるツイートを使用して，市民の移動を分析する．まず，2.1 節で，SNS を使用した観光や移動についての研究を紹介する．次に，2.2 節で単語の分散表現を活用した研究を紹介し，本研究の位置づけについて述べる．

2.1 SNS を利用した観光や移動の分析

SNS の発達により，観光客の情報発信が容易になっており，観光地としての発展や地域振興には，SNS の情報収集が必要である．松本ら [25] は，Twitter から地域情報や観光情報を，共起する言葉に着目して収集した．まず，1 年分のツイートを集めた長期データと，1 日分のツイートを集めた短期データを作成した．この研究では，インターネット上に存在する膨大なデータの中から，地域や観光に役立つデータを抽出することを目標にしている．結果として，1 年分の長期データには，「花火大会」といった季節ごとの観光情報が含まれており，1 日分の短期データには，その 1 日が含まれる時期の情報や，「フグ」のような季節を問わない情報が含まれていた．長島ら [23] は，Twitter を用いて，地域ユーザの口コミを収集した．まず，店舗について言及している口コミツイートを，ツイート中の地域名と，ツイートに付随する位置情報を用いて収集した．

本研究では，長島らの研究と同様にプロフィール情報を用いて居住地を推定する．また，観光や地域に関する情報の収集には，共起語に着目した研究が多くみられるが，ここでは，通勤や通学で訪れる目的地と，旅行で訪れる観光地では，ツイートの内容が大きく異なると考え，前後の文脈を考慮して生成される単語の分散表現を用いる手法を提案する．

井上ら [12] は，位置情報付きツイートを解析し，人々の行動に関する情報を自動抽出した．まず，収集した位置情報付きツイートの中から，秋葉原駅を中心に 500m 四方の領域で発言されたツイートを抽出した．秋葉原駅を選んだ理由は，来訪者の目的が比較的分かりやすいためとされている．次に，収集したツイートを，人手で「食」や「買い物」などのカテゴリに分類した．さらに，人手で分類したツイートを元に，カテゴリが未

知のツイートの自動分類を行った．この際，4,558 単語の出現頻度が，カテゴリごとに異なることを利用している．本研究の提案手法では，特定の都市の市民の認識に着目し，「秋葉原駅」のような特定のスポットに限定はしない．また，多くの地名を対象として，地名と移動目的を表す単語との文脈的な類似性を手がかりとして，市民の認識に基づく地名の可視化を行う．

2.2 単語の分散表現を活用した研究

丸井ら [26] は，SNS においてコミュニティごとで意味の異なる単語を単語の分散表現によって明らかにした．まず，Twitter を用いて，ユーザ同士の会話からコミュニティを抽出し，さらに，属するユーザのプロフィールからコミュニティのラベルを付けた．次に，コミュニティごとに意味の違いを把握するため，ユーザのツイートをコミュニティごとにまとめたコーパスで skip-gram を学習させた．そして，単語の分散表現に基づき，コミュニティごとで使われ方の異なる単語を，いくつかの例を用いて検証した．

この研究では，ユーザのコミュニティごとで意味の異なる単語について，単語の分散表現を用いることで獲得する手法を提案した．本研究では，都市ごとに単語の分散表現を学習させることにより，同じ単語でも都市が異なれば，市民による認識が異なることに着目する．したがって，目的は異なるが，ユーザのコミュニティごとに単語の分散表現を学習させる点と，コミュニティごとに意味が異なる単語に着目する点は共通している．

開地ら [14] は，単語の分散表現を使用してユーザの潜在的興味を発見し，その潜在的興味に基づく観光地推薦システムを作成した．まず，Wikipedia(注3) と Yahoo!知恵袋 (注4) を用い，観光地を集めた観光地データベースを作成する．そして，Yahoo! 知恵袋の国内旅行に関連のあるカテゴリについて，観光地データベース内の観光地を検索クエリとして投稿を収集し，コーパスを作成する．次に，skip-gram を用い，コーパスを学習させて，観光地ごとの単語の分散表現を獲得する．この研究では，ユーザに好きな観光地と嫌いな観光地を入力させて，単語の分散表現による観光地の類似度に基づき，観光地を推薦している．

さらに開地ら [15] は，観光地データベースをより洗練されたデータベースにするために，Google Place API(注5) を用いて観光地を増やして，人手で評価実験を行っている．これらの研究により，意味的に関連の強い観光地は，単語の分散表現も近くなることを可視化により示している．

本研究でも，単語の分散表現を利用するが，地名同士の類似度を計算するのではなく，地名と認識を表す単語との類似度に着目する点で異なる．さらに本研究では，複数の都市について単語の分散表現を生成することで，都市ごとの市民の地名に対する認識の違いを明らかにする．

3. 都市を単位とした Twitter ユーザの収集

本章では，都市ごとに市民の移動を分析する方法を提案する．提案手法の概要を図 2 に示す．

(注3)：https://ja.wikipedia.org/wiki
(注4)：https://chiebukuro.yahoo.co.jp/
(注5)：https://developers.google.com/places/

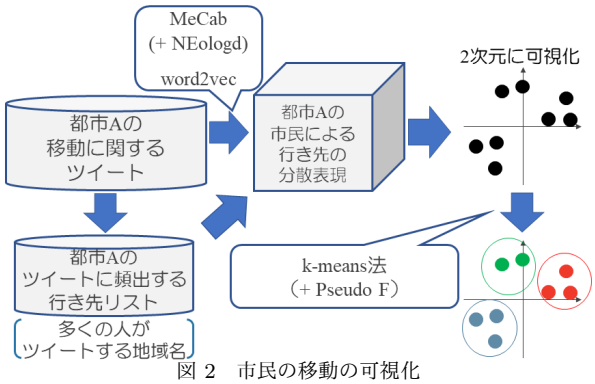

図 2　市民の移動の可視化

本研究は，まず，都市ごとに Twitter ユーザを収集し，市民の移動に関するツイートを収集する．次に，ツイートに頻出する地名を選択し，行き先リストを作成した後，行き先リスト中の地名の分散表現に基づいて 2 次元に可視化を行う．最後に，2 次元に可視化された行き先について，クラスタリングを行い，市民の認識が似ている行き先をまとめる．

3.1 都市を単位とした Twitter ユーザの収集

本研究では，長島ら [23] の手法を参考に，居住地域ごとに Twitter ユーザを収集する．

まず，ツイプロ (注6) を使用し，対象とする市区町村を居住地としてプロフィールに記述しているユーザの収集を行う．しかし，ツイプロには登録されていないユーザもいるため，位置情報に市区町村を記載しているすべての Twitter ユーザを収集することはできない．次に，ツイプロを使用して収集したユーザが，10 人以上フォローしているユーザを拡張し，プロフィールの位置情報に，対象とする市区町村を含むユーザを追加する．

3.2 市民の移動に関するツイートを収集

3.1 節で収集したユーザのツイートから，移動に関するツイートを収集する．まず，移動に関係がある動詞を抽出する．移動に関係がある動詞のリストは，3.1 節で収集したユーザのツイートを用いて，skip-gram で「行く」の各活用形のコサイン類似度を計算し，上位 4 単語について，さらに活用形を展開したもの追加する．移動に関するツイートは，地名の後 10 形態素以内に移動に関する動詞，または「旅行」が出現しているツイートとしている．移動に関するツイートの例を図 3 に示す．

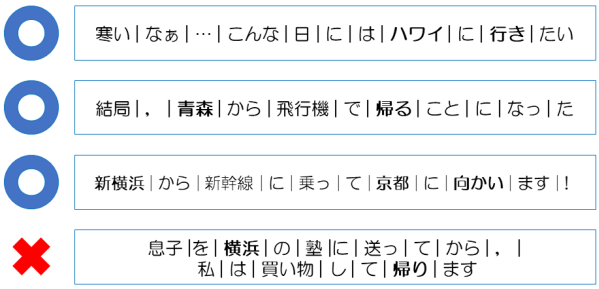

図 3　移動に関するツイートの例

なお，品詞や地名の判定には，形態素解析器として MeCab [5]，

(注6)：https://twpro.jp/

辞書として NEologd [17] を使用する．さらに，bot や宣伝と思われるツイートを除外している．また，分散表現を生成する過程で，前後の文脈として悪影響を与えるツイートの要素は以下の方法で除外している．

- 「I'm at」に続く部分
- 「#」に続く部分

3.3 行き先リストの作成

都市ごとに行き先リストを作成する．具体的には，3.2 節で収集した移動に関するツイートから，「つくばから」といった移動の出発点を表している地名を除外して，投稿人数を計算した．さらに，地名として不適切な「動物園」などは人手で除去し，投稿人数の上位 20 件を行き先リストとする．以上により，多くの人がツイートする地名が，行き先リストとして獲得できる．

寒い | なぁ | … | こんな | 日 | に | は | ハワイ | に | 行き | たい

結局 | ， | ~~青森~~ | から | 飛行機 | で | 帰る | こと | に | なっ | た

~~新横浜~~ | から | 新幹線 | に | 乗っ | て | 京都 | に | 向かい | ます | ！

図 4　行き先リスト作成の例

3.4 行き先リスト中の地名の分散表現を利用した可視化

まず，3.2 節で収集したツイートを用い学習を行い，3.3 節で作成した行き先リスト中の地名の分散表現を生成する．分散表現の学習には skip-gram を使用する．また，生成した分散表現に基づき，以下の手順で，2 次元での可視化を行う．なお，以下の手順では，x 軸の可視化に用いる市民の認識を表す語を $keyword_{x+}$，$keyword_{x-}$，y 軸の可視化に用いる市民の認識を表す語を $keyword_{y+}$，$keyword_{y-}$ とする．

（1） 3.2 節で収集したツイートを用い，$keyword_{x+}$ と $keyword_{x-}$ の分散表現を獲得

（2） $keyword_{x+}$ と 3.3 節で獲得した行き先リスト中の地名とのコサイン類似度を計算し，Z スコアを用いて正規化する

（3） $keyword_{x-}$ に対し，2 と同様の処理を行う

（4） 地名ごとに，2 で正規化された $keyword_{x+}$ とのコサイン類似度から，3 で正規化された $keyword_{x-}$ とのコサイン類似度の差分を計算し，これを x 座標とする

（5） $keyword_{y+}$，$keyword_{y-}$ について，1〜4 と同様の処理を行い，これを y 座標とする

例えば，$keyword_{x+}$ が動詞「行く」，$keyword_{x-}$ が動詞「帰る」，$keyword_{y+}$ が希望を表す助動詞「たい」，$keyword_{y-}$ が過去・完了・存続を表す助動詞「た」の場合，将来行きたい地名なのか，過去に行った地名なのかを区別できるように，地名を 2 次元平面上にプロットし，可視化を実現する．

3.5 地名のクラスタリング

都市ごとの市民の行き先に対する認識の傾向を明らかにするため，3.4 節で計算した座標を利用して，クラスタリングを行う．クラスタリングは k-means 法 [7] により分類する．k-means 法は，非階層的クラスター分析の代表的な手法で，各クラスタの重心とクラスタとの距離の和が最小になるよう，与えられたクラスタ数 k 個に分類する手法である．クラスタ数の決定には，Pseudo F [2] を用いている．Pseudo F は，クラスタ間の分散と，クラスタ内の分散を考慮した指標で，式 (1) において Pseudo F が最大となったクラスタ数 k を用いてクラスタリングを行う．

$$PseudoF = \frac{\text{クラスタ間距離二乗和}/(\text{クラスタ数 } k-1)}{\text{クラスタ内距離二乗和}/(\text{全サンプル数} - \text{クラスタ数 } k)} \quad (1)$$

4. 実験：市民の移動の傾向の分析

4.1 目　　的

本章では，3 章で述べた提案手法の有効性を検証するために，複数の都市を対象として，市民の移動の傾向の分析を行う．移動には，「行く」と「帰る」という正反対な 2 つの向きが存在する．さらに，過去に完了したことを示す「行った（帰った）」と，将来への願望を表す「行きたい（帰りたい）」という正反対の 2 つの認識が存在する．本実験では，行き先を表す地名に関して，「行く場所か，帰る場所か」，「行きたい（帰りたい）場所か，実際に行った（帰った）場所か」を調査することで，市民の認識を表す 2 つの単語と行き先の地名との単語の分散表現の類似度に基づく可視化の妥当性に関して検証を行う．また，可視化された地名をクラスタリングすることにより，市民が同じ認識を持つ地名が正しく分類できているかを検証する．

4.2 実験データ

実験データは，北海道，茨城県，南関東，福岡県にある 8 つの都市，すなわち，札幌市，水戸市，つくば市，千葉市，さいたま市，横浜市，北九州市，福岡市に居住しているユーザが，平成 29 年 7 月 1 日から 11 月 30 日までの 5ヶ月間に投稿したツイートを対象とする．また，これらのツイートを用いて，動詞リスト，移動に関するツイート，行き先リストを作成する．移動に関するツイートの収集結果を表 1 に示す．

表 1　発見したユーザの評価値の比較

都市名	ツイート数
札幌市	15,712
つくば市	18,615
水戸市	11,136
横浜市	28,399
千葉市	5,563
さいたま市	12,800
北九州市	6,162
福岡市	10,120

4.3 実 験 方 法

本実験では，市民の移動について，市民が「行く場所か帰る場所か」，「行きたい（帰りたい）場所か，実際に行った（帰っ

た）場所か」を明らかにするため，行き先リスト中の地名と，動詞「行く」と動詞「帰る」，希望を表す助動詞「たい」と過去・完了を表す助動詞「た」とのコサイン類似度に基づいて地名を可視化する．可視化された地名について，国勢調査等を用いて第一著者が評価する．評価は，表 2 に従い，点数を付ける．また，20 件の地名についてすべて 2 点と評価した場合を 100 点満点とし，都市ごとに採点を行う．

表 2　評 価 方 法

点数	0	1	2
評価	理解が難しい	どちらともいえない	理解できる

さらに，市民が同じ認識を持つ地名を分類するため，可視化された地名をクラスタリングする．可視化された地名のクラスタリングについて，Ly ら [6] が用いた手法を参考に，式 (2) に示す純度を用いてクラスタリング結果を評価する．

$$\text{純度} = \sum_i \frac{|C_i|}{n} \max \text{精度}\,(C_i, L_j) \qquad (2)$$

C は，提案手法により分類されたクラスタ，L は，評価者が人手で分類したクラスタを示している．n は，行き先リストの地名数なので，本研究では $n = 20$ となっている．また，提案手法では C_i と分類し，評価者が L_j と分類した精度を式 (3) に示す．

$$\text{精度}\,(C_i, L_j) = \frac{|C_i \cup L_j|}{|C_i|} \qquad (3)$$

本実験において評価者は，実験を行った都市，または隣接する都市に居住経験のある人とする．なお，クラスタ数は，提案手法によるクラスタリング結果と同数とした．

4.4　実 験 結 果

札幌市，水戸市，つくば市，千葉市，さいたま市，横浜市，北九州市，福岡市の結果をそれぞれ図 5，図 6，図 7，図 8，図 9，図 10，図 11，図 12 に示す．x 軸は，「行く」と「帰る」といった移動の向きを表しており，$0 < x$ のとき，市民が「行く」地名であり，$0 > x$ のとき，市民が「帰る」地名であることを示している．同様にして，y 軸は，「たい」と「た」といった市民の認識を表しており，$0 < y$ のとき，市民が過去に「行きたい（帰りたい）」という願望もつ地名であり，$0 > y$ のとき，市民が過去に「行った（帰った）」地名であることを示している．また，各クラスタは，異なる色で示されている．

著者が表 2 の基準に基づき，各都市における地名の可視化について採点結果を表 3 に示す．最後に，提案手法によるクラスタリングの結果を純度に基づき評価した結果を，表 4 に示す．

表 3　可視化の採点結果

都市名	札幌	水戸	つくば	千葉	さいたま	横浜	北九州	福岡	平均
点数	75.0	52.5	72.5	55.0	55.0	75.0	72.5	80.0	67.2

表 4　地名クラスタの純度

都市名	札幌	水戸	つくば	千葉	さいたま	横浜	北九州	福岡	平均
純度	0.65	0.55	0.65	0.65	0.75	0.70	0.70	0.60	0.66

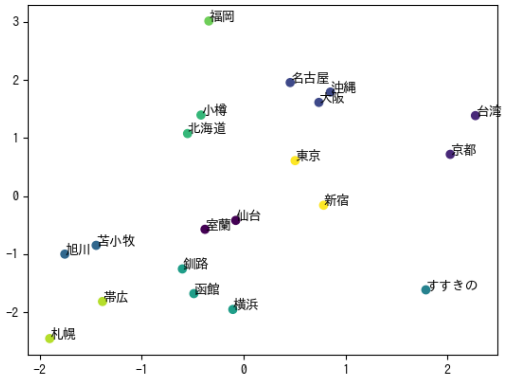

図 5　札幌市での実験結果

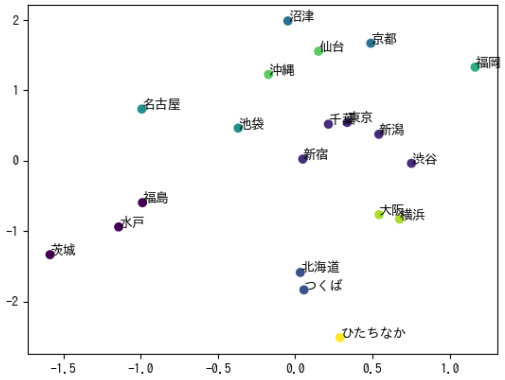

図 6　水戸市での実験結果

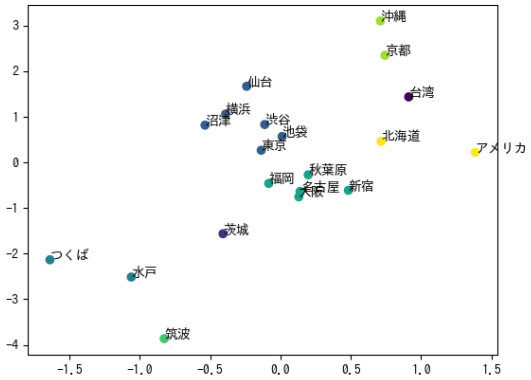

図 7　つくば市での実験結果

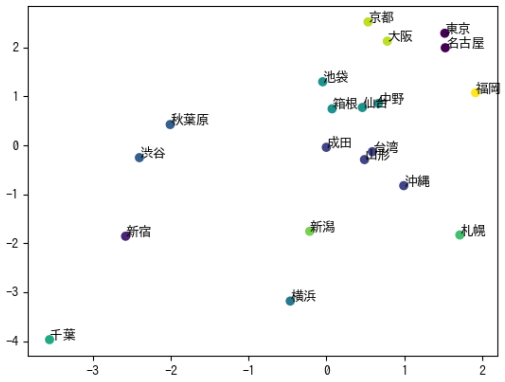

図 8　千葉市での実験結果

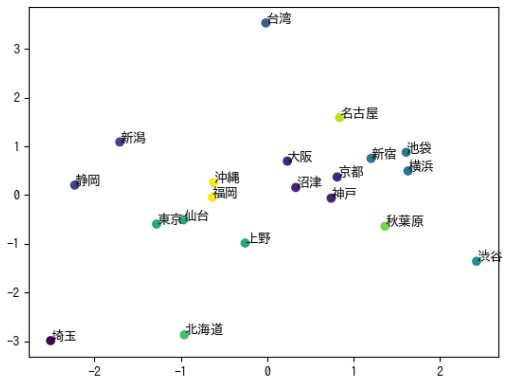

図 9　さいたま市での実験結果

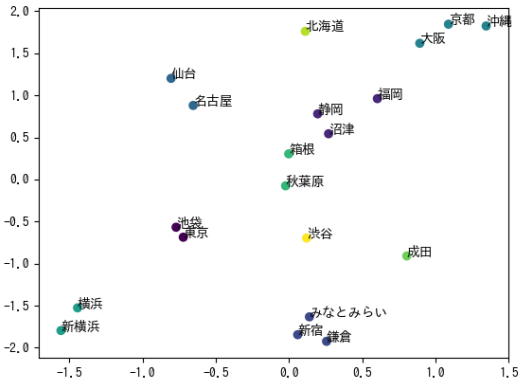

図 10　横浜市での実験結果

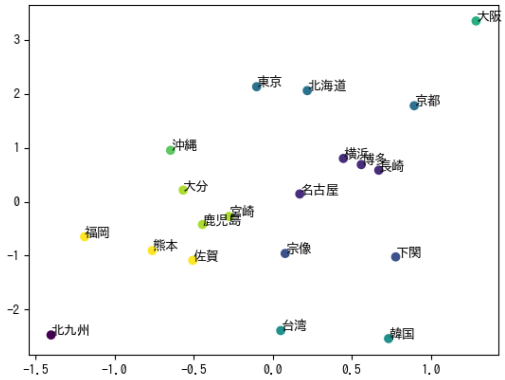

図 11　北九州市での実験結果

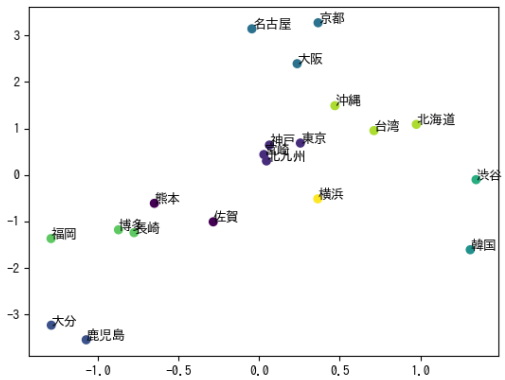

図 12　福岡市での実験結果

5.　考　　察

本章では，4.4 節の実験の結果について考察する．5.1 節では，分散表現の類似度に基づく地名の可視化について考察する．5.2 節では，可視化した地名のクラスタリングについて考察する．5.3 節では，季節による行き先の変化について考察する．

5.1　市民コーパスを利用した地名の可視化による考察

4.4 節の結果をふまえて，市民の認識を表す語と地名との単語の分散表現の類似度の可視化に関して検証する．5.1.1 節では，すべての都市における，観光地，居住地，帰省地の可視化について考察する．5.1.2 節では，著者による採点結果が 70 点以上であった札幌市，つくば市，横浜市，北九州市，福岡市について考察する．5.1.3 節では，採点結果が 70 点未満であった水戸市，千葉市，さいたま市について考察する．

5.1.1　観光地，居住地，帰省地の可視化

- 全ての都市に共通して見られる特徴として，京都に代表される観光地は，第 1 象限（行きたい）に可視化されている．平成 27 年の共通基準による観光入込客統計 [30] によると，府外から京都を訪れた観光客は東京都に次いで 2 位である．また，2015 年，日本政府観光局 [33] によると，京都は，アメリカの旅行雑誌最大手の「Travel+Leisure」誌の，世界の魅力的な都市ランキングにおいて，京都が 2 年連続 1 位であると公表している．以上のことから，京都が第 1 象限（行きたい）に可視化されていることは妥当である．京都の他にも，第 1 象限には，東京，名古屋，大阪など，観光客が多い東海道新幹線の沿線にある大都市や，沖縄や北海道といった人気の観光地を確認することができる．これらの都市には，「行きたい」と考えている市民が多くいるので，「行きたい」地域としては妥当である．
- 居住地は，原点から遠い位置の第 3 象限（帰った）に可視化されている．これは，居住地から別の都市に移動した市民が，帰る際，または帰ってきてからツイートすることが多いためであり，実際のツイートからも確認することができ，妥当である．他にも，水戸市の結果における「茨城」や，つくば市民の結果における「筑波」のように，居住している都市が属する地域名や，別名が可視化されている点は，実際のツイートからも居住している都市名と同様の使い方をしているものが多く見受けられ，妥当である．また，横浜市の結果における「新横浜」や，福岡市における「博多」は，新幹線等の交通機関を使い，居住している地域に帰ってきたことを表しており，妥当である．
- 札幌市，北九州市，福岡市の結果では，第 3 象限（帰った）に北海道内，九州内の地名を多く確認することができる．これは，帰省をしたことを表すツイートが多く収集されていたことによるものである．国勢調査 [29] の結果からも，札幌市には北海道内から，北九州市と福岡市には九州内から多くの市民が転入していることがうかがえる．

5.1.2　採点結果が 70 点以上である都市についての考察

著者による採点結果が 70 点以上であった札幌市，つくば市，横浜市，北九州市，福岡市について考察する．

- 札幌市について

図 5 において，道内の各都市は，「札幌」と近い結果となってい

る．このことは，国勢調査 [29] によると，道内から札幌市に転入した市民は，道外から転入した市民の比べて多いことや，大多数の市民が市内，または道内の都市に通勤・通学していることを示していることから妥当である．

- つくば市について

図 7 においては，Twitter を盛んに使用する筑波大学の学生の特徴が表れている．具体的には,「つくば」に帰りたいとツイートする市民が多いことである．これは，筑波大学の学生が，地元に帰省した際などに,「つくば」に帰ってきたいという意図を示している．また,「沖縄」,「北海道」,「台湾」,「アメリカ」といった，学生に人気がある地域に行きたいと考える市民が多い傾向も確認できる．

- 横浜市について

図 10 においては，横浜市から比較的行きやすい地名が多くみられる.「みなとみらい」,「新横浜」,「鎌倉」といった，横浜市内や横浜市近郊の地域には，行ったり帰ったりしている様子がうかがえる．また,「池袋」,「渋谷」,「静岡」といった，横浜市からは比較的近い地名が $y = 0$ 付近にみられるが，実際のツイートからも，過去に行ったことを表すものと，行きたいという願望を表すものとの両方を確認できる．

- 北九州市，福岡市について

図 11 と図 12 から，2 つの都市には共通して，九州内の各都市に「帰る」傾向があるが，実際に，北九州市と福岡市には，九州内の各県から多くの人が転入している．また,「韓国」には「行った」傾向が確認できる．福岡県から韓国には短時間で行くことができ，実際のツイートからも韓国に行った旨を示すものが多くみられた．

5.1.3 採点結果が 70 点未満である都市についての考察

著者による採点結果が 70 点未満であった水戸市，千葉市，さいたま市について考察する．

- 水戸市について

図 6 において，水戸市の特徴的な点は,「ひたちなか」に行った市民が多いことである．ツイートからは，国営ひたち海浜公園や，ROCK IN JAPAN FESTIVAL(注8)に行ったことを報告するものが多くみられた．一方で，新潟が第 1 象限（行きたい）に可視化されている点は，不適当な結果といえる．これは,「新潟鳥栖札幌は行きたいね」といった，新潟をホームとするサッカーチームを指しているツイートが多いことが原因である．

- 千葉市について

図 8 においては，鉄道に関するツイートにより，理解できない結果になった地名があった．例えば,「中野」は，中野駅が鉄道の終点であることを示す「中野行き」を意味するツイートが多く見受けられた．千葉市が良くない結果になった理由は，ツイート数が少ない千葉市では特に，市民の移動に関係のないツイートが悪影響を与えたためである．

- さいたま市について

図 9 においては，交通の便が良い東京都内の地名が $x > 0$ に多い点は理解できる．実際のツイートからも，特に「池袋」や「渋谷」を訪れるツイートが多く確認できた．一方で，観光地として人気がある北海道や福岡や沖縄が正しく可視化できていない．これは，東京都内の地名を含むツイートに比べ，北海道や福岡や沖縄に関するツイートが少なかったことが原因と考える．

5.2 クラスタリング結果について考察

評価者の人手によるクラスタリングに基づいて考察する．純度は，平均で 0.66 となっている．まず，評価者によるクラスタリングと，提案手法によるクラスタリングで分類が一致した地名について考察する.「北海道」,「茨城」,「千葉」など，都市名を包含する地名（県名）は，提案手法と評価者によるクラスタリングのいずれでも，1 つだけの要素のクラスタに分類された．また，観光目的など，その地域に向かう大多数の市民ユーザの目的が共通している場合，精度よくクラスタリングができている．例えば，北九州市における「韓国」と「台湾」があげられる．一方で，原点近くに可視化された地名，具体的には，東京都内の地名や，大阪や名古屋などの大都市は，評価者によるクラスタリングと異なる結果となる傾向が見られた．

次に，純度が高い都市と低い都市について考察する．さいたま市は，距離の離れた地名が提案手法により適切にクラスタリングされており，評価者の結果と一致したことから，純度が高い結果となった．一方，水戸市では，距離の離れた地名が提案手法により適切にクラスタリングされていないことから，純度が低い．また，福岡市は，九州内の帰省地にあたる地名を評価者が同じクラスタに分類したのに対して，提案手法では異なるクラスタに分類していたことから，純度が低い結果となった．

5.3 季節による行き先の変化についての考察

都市ごとの行き先は，季節により変化する．本手法では，データの収集期間を変更することで，季節に応じた行き先となる地名の可視化も季節に応じたものになると仮定する．そこで，季節ごとのデータを用いることにより，季節による行き先の変化を確認する．2017 年 6 月から 2017 年 8 月の 3ヶ月間を夏，2017 年 11 月から 2018 年 1 月の 3ヶ月間を冬として，行き先を可視化する実験を行った．夏と冬とを比較した結果，水戸市では，8 月に音楽フェスである ROCK IN JAPAN FESTIVAL が開催される「ひたちなか」が，夏に「行った」側に可視化されているのに対し，冬には「帰った」側に可視化されている．また，札幌市では，8 月に音楽フェスである活性の火 (注9)が開催される「苫小牧」が，夏に「行きたい」側に可視化されているのに対し，冬には「帰りたい」側に可視化されている．この結果から，収集期間の変更により，季節に応じた行き先を可視化できることを明らかにできた．

6. おわりに

本稿では，ツイートを使用して，都市ごとに市民の移動の傾向を分析する手法を提案した．8 つの都市ごとに収集した Twitter ユーザについて，5ヶ月分のツイートを用いて，提案手法の有効性を検証するために，市民の行動や移動に関する認識を分析する実験を行った．実験の結果，人気のある観光地が

(注8)：http://hitachikaihin.jp/

(注9)：https://activefire14.jimdo.com/

「行きたい」地名として可視化され，居住している都市名や帰省先の地名が「帰った」地名として可視化されたことで，単語の分散表現のコサイン類似度に基づく可視化の有効性を示した．また，クラスタリングについては，純度が平均 0.66 となっており，観光地や居住地など，行き先としての目的が明確な地名を分類できることを示した．今後の課題として，「#聖地巡礼」等のハッシュタグに着目することで，聖地巡礼を目的とする観光地域の抽出についても検討している．

謝辞　本研究の一部は，科学研究費補助金基盤研究 B（課題番号 16H02913）の助成を受けて遂行された．

文　献

[1] Bojanowski, Piotr; Grave, Edouard; Joulin, Armand; Mikolov, Tomas; “Enriching Word Vectors with Subword Information”. Transactions of the Association for Computational Linguistics. Vol. 5, 2017, p. 135-146.

[2] Calinski, Tadeusz; Harabasz, Joachim. “A Dendrite Method for Cluster Analysis”. Communications in Statistics - Theory and Methods. US, Marcel Dekker Inc., 1974, p. 1-27.

[3] Firth, John Rupert. “A Synopsis of Linguistic Theory 1930-55.”. Selected Papers of J. R. Firth 1952-59. Robert, Palmer Frank, ed. London, Longmans, 1957, p. 168-205.

[4] Harris, Zellig Sabbettai. “Distributional Structure”. Papers on Syntax. Dordrecht, Nederland, Springer, 1954, p. 146-162.

[5] Kudo, Taku; Yamamoto, Kaoru; Matsumoto, Yuji. “Applying Conditional Random Fields to Japanese Morphological Analysis”. Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP2004). Barcelona, Spain, 2014, p. 230-237.

[6] Ly, Duy Khang; Sugiyama, Kazunari; Lin, Ziheng; Kan, Min-Yen. “Product Review Summarization from a Deeper Perspective”. Proceedings of the 11th annual international ACM/IEEE Joint Conference on Digital Libraries (JCDL2011). New York, USA, ACM, 2011, p. 311-314.

[7] MacQueen, James. “Some Methods for Classification and Analysis of Multivariate Observations”. Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability. Berkeley, California, USA, 1967, p. 281-297.

[8] Mikolov, Tomas; Chen, Kai; Corrado, Greg; Dean, Jeffrey. “Efficient Estimation of Word Representations in Vector Space”. Proceedings of the International Conference on Learning Representations (ICLR2013). Scottsdale, Arizona, USA, 2013.

[9] Mikolov, Tomas; Sutskever, Ilya; Chen, Kai; Corrado, Greg; Dean, Jeff. “Distributed Representations of Words and Phrases and their Compositionality”. Proceedings of the Neural Information Processing Systems 2013 (NIPS2013). Stateline, Nevada, USA, 2013.

[10] 荒堀 智彦. “2009 年新型インフルエンザ A(H1N1)pdm の流行とローカルな伝播過程”. 日本地理学会発表要旨集. 公益社団法人 日本地理学会, 2014, vol. 2014a, p. 100114.

[11] 井川洋平. “ソーシャルメディア位置情報分析を行う前の Tips”. SlideShares. 2017-12-05. https://www.slideshare.net/YoheiIkawa/tips-55961281, (参照 2017-12-05).

[12] 井上拓也, 山田剛一, 増田英孝, 荒牧英治, 中川裕志. “ソーシャルメディア上の位置情報付きテキストを利用した行動分析”. 研究報告情報基礎とアクセス技術（IFAT）. 情報処理学会 情報基礎とアクセス技術研究会, 2013, vol. 2013-DD-89, no. 2, p. 1-7.

[13] 岩波データサイエンス刊行委員会編. 岩波データサイエンス Vol.2. 岩波書店, 2016, 152p.

[14] 開地亮太, 檜垣泰彦. “潜在的興味に基づく観光地推薦システムの試作”. 電子情報通信学会技術研究報告. 電子情報通信学会, 2015, vol. 115, no. 138, p. 29-34.

[15] 開地亮太, 檜垣泰彦. “単語の分散表現を使用した観光地推薦システムの構築”. 電子情報通信学会技術研究報告. 電子情報通信学会, 2016, vol. 115, no. 486, p. 45-50.

[16] 神嶌敏弘編. 深層学習. 近代科学社, 2015, 267p.

[17] 佐藤敏紀, 橋本泰一, 奥村学. “単語分かち書き辞書 mecab-ipadic-NEologd の実装と情報検索における効果的な使用方法の検討”. 言語処理学会 第 23 回年次大会 発表論文集. 言語処理学会, 2017, p. 875-878.

[18] 杉浦聡志, 町勉, 塚本圭, 高木朗義, 倉内文孝. “道路統廃合を念頭にした生活道路ネットワークデザインモデルの実装に向けた拡張”. 土木学会論文集 F4（建設マネジメント）. 公益社団法人 土木学会, 2015, vol. 71, no. 4, p. I_53-I_63.

[19] 辰巳浩, 堤香代子, 吉城秀治, 鶴丸梓. “世帯属性や移動環境を考慮した地域公共交通の需要予測に関する研究”. 交通工学論文集. 一般社団法人 交通工学研究会, 2016, p. A_100-A_107.

[20] 坪井祐太, 海野裕也, 鈴木潤著. 深層学習による自然言語処理. 講談社, 2017, 229p.

[21] ダヌシカ ボレガラ, 岡崎直観, 前原貴憲. ウェブデータの機械学習. 講談社, 2016, 186p.

[22] 中澤高志. “職業別純移動による東京圏の居住地域構造”. 日本地理学会発表要旨集. 公益社団法人 日本地理学会, 2015, vol. 2015s, p. 100074.

[23] 長島里奈, 関洋平, 猪圭. “地域ユーザに着目した口コミツイート収集手法の提案”. 第 8 回データ工学と情報マネジメントに関するフォーラム（DEIM2016）, データ工学と情報マネジメントに関するフォーラム, 2016, B4-3.

[24] 藤田翔平. “東京大都市圏郊外地域における就業核の成長”. 日本地理学会発表要旨集. 公益社団法人 日本地理学会, 2015, vol. 2015s, p. 100090.

[25] 松本義之, 藪内賢之. “Twitter からの地域・観光情報収集とその有用性の検討”. 第 27 回バイオメディカル・ファジィ・システム学会年次大会 講演論文集. バイオメディカル・ファジィ・システム学会, 2014, p. 87-88.

[26] 丸井淳己, 則のぞみ, 榊剛史, 森純一郎. “分散表現を用いたコミュニティにおける単語使用傾向の分析”. 人工知能学会第 28 回全国大会. 人工知能学会, 2014, 4I1-3.

[27] 山口裕通, 奥村誠, Tirtom, Huseyin. “都市間交通需要の LOS 弾力性に関する研究”. 土木学会論文集 D3（土木計画学）. Japan Society of Civil Engineers, 2013, vol. 69, no. 5, p. I_629-I_638.

[28] 吉田朋史, 北山大輔, 中島伸介, 角谷和俊. “ユーザレビュの分散表現を用いた主観的特徴の意味演算による観光スポット検索システム”. 第 9 回データ工学と情報マネジメントに関するフォーラム（DEIM2017）, データ工学と情報マネジメントに関するフォーラム, 2017, P6-5.

[29] 「平成 27 年国勢調査」. 移動人口の男女・年齢等集計（人口の転出入状況）. 2017 年 1 月 27 日公表.

[30] 「共通基準による観光入込客統計」. 観光庁. 2017 年 11 月 30 日公表.

[31] “SNS 等を活用した分析”. 観光庁. http://www.mlit.go.jp/common/001086249.pdf, (参照 2017-11-23).

[32] “平成 27 年度 ICT を活用した訪日外国人観光客動向調査事業実施報告書”. 観光庁. http://www.mlit.go.jp/common/001158957.pdf, (参照 2017-11-23).

[33] “米大手旅行雑誌「Travel+Leisure」誌観光ランキングで京都が 2 年連続世界一に”. 日本政府観光局. https://www.jnto.go.jp/jpn/news/press_releases/pdf/20150708.pdf, (参照 2017-12-19).