位置情報による分散表現を用いたユーザの移動の分析

野島　僚太† 　廣田　雅春† 　石川　博††

† 岡山理科大学総合情報学部情報科学科　〒 700-0005　岡山県岡山市北区理大町 1-1
†† 首都大学東京大学院システムデザイン研究科　〒 191-0065 東京都日野市旭が丘 6-6
E-mail: †i15i054nr@ous.jp, ††hirota@mis.ous.ac.jp, †††ishikawa-hiroshi@tmu.ac.jp

あらまし　近年，多くの人々がソーシャルメディアサイトに緯度経度情報が付与されたコンテンツを投稿している．観光地におけるこのようなデータは，観光ルートや観光スポットの推薦などの研究に利用されている．本研究では，これらの研究に用いることが可能な移動による分散表現を作成するための手法を提案する．提案手法では，緯度経度情報から表される 2 点間のユーザの移動について，Skip-gram モデルに基づくアルゴリズムにより学習することで，移動に関する分散表現を作成する．Skip-gram モデルでは，文書中のある単語に対して，その前後の単語を同じ文脈として学習するが，本研究では，Skip-gram モデルを改良し，ユーザの移動軌跡中のある地点から移動先のみを対象として学習する．作成したベクトルは，Word2vec と同様に意味演算を行うことが可能であるため，意味演算の結果について分析と考察を行う．

キーワード　Skip-gram model，Twitter，地理情報，移動軌跡

1 はじめに

近年，観光産業は，日本において重要な地位を占めており，多くの国内旅行者が存在する．国土交通省観光庁の旅行・観光消費動向調査 [1]によると，2017 年の日本人国内延べ旅行者数は，約 6 億 4,720 万人で，前年比約 1.0%増となっている．また，スマートフォンなどのデジタルデバイスの普及に伴い，多くの人々が Twitter[2]や，Flickr[3]などのソーシャルメディアサイトにユーザの自身の状況や，飲食店や観光スポットの感想などを投稿している．また，これらのコンテンツには，Global Positioning System (GPS) が測位した緯度経度情報が付与されていることが多い．そのため，投稿された地点の緯度経度情報や時間などの情報から人々の移動の分析を行うことで観光分野への応用が可能となる．

ソーシャルメディアを用いて観光の行動の分析や，観光スポットの推薦などの研究が行われている．その中で，観光客の移動に関する分析 [14] や観光ルートの推薦 [9] など観光客の移動に関する研究が行われている．本研究では，観光地での人々の移動を分析するために，移動による分散表現を作成するための手法を提案する．提案手法では，ソーシャルメディアサイトに投稿されたユーザのコンテンツに付与された緯度経度からなる移動軌跡に対して，Word2vec に基づくアルゴリズムを適用することである地域におけるユーザの 2 点間の移動に基づく分散表現を作成する．

Word2vec とは，Mikolov ら [5] によって提案されたニューラルネットワークを用いた単語の出現する順番に基づいた単語のベクトルを学習する手法である．この手法は，類似した文脈で用いられる単語は，類似した意味を持つという分布仮説に基づいている．そして，Word2vec では，学習する手法として，Continuous Bag of Words (CBOW) モデルと Skip-gram モデルの二つのモデルが提案されている．ここで，本研究で用いる Skip-gram モデルについて簡単に説明する．図 1 に Skip-gram モデルを示す．Skip-gram モデルでは，文書中のある単語 $w(t)$ から，その周辺の単語 $w(t-c),\ldots,w(t-1),w(t+1),\ldots,w(t+c)$ を予測するタスクをニューラルネットワークで学習する．ここで，c はウインドウサイズと呼ばれるパラメータであり，学習に用いる前後のそれぞれの単語数である．この単語 $w(t)$ と周辺の単語の関係をニューラルネットワークで学習し，学習後の中間層の各単語の重みを抽出することにより，各単語のベクトルを得ることができる．このモデルによって構築された単語のベクトル空間において，ある単語の周辺によく現れる単語同士は類似度が高くなる．これを利用して，単語間の類似度の計算や単語の持つ意味の加算，減算が可能になる．

本研究では，ソーシャルメディアサイトのコンテンツに付与されている緯度経度情報から得られたユーザの移動を学習することにより，移動に関する分散表現を作成する手法を提案する．このとき，ユーザの 2 点間の移動を対象として学習を行うことで分散表現を作成する．Skip-gram モデルは，学習を行う単語の前後の単語を同じ文脈として学習を行うため，本研究で扱う移動先の地域のみを学習させるという点では，従来の Skip-gram モデルをそのまま適用することはできない．

そこで，本論文では，Skip-gram モデルを改良し，ある地域から次の地域への移動のみを学習するようにした Skip-gram モデルを用いて移動軌跡からユーザの移動を学習し，ある地域についてベクトルを作成する手法を提案する．

また，GPS による誤差や緯度経度情報の桁数などの差に考慮しつつ学習を簡略化するために，緯度と経度を用いて直接学習するのではなく，分析する領域全体を小さい領域を表すセル

1 : http://www.mlit.go.jp/kankocho/siryou/toukei/shouhidoukou.html
2 : https://twitter.com/
3 : https://www.flickr.com/

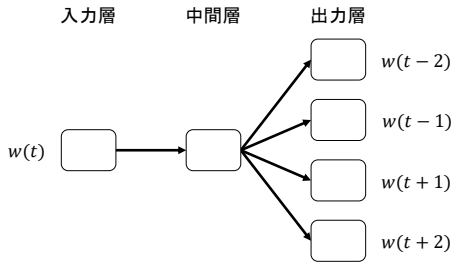


図 1: Skip-gram モデル

への量子化を行った上で学習する．本研究では，四角形のセルを用いたが，領域に重複がなければ，どのような形状でも構わない．

また，Word2vec では，文書中のある単語から周辺の単語を学習することでそれぞれの単語の特徴をベクトルとして表現するが，本研究では，文の代わりに，ソーシャルメディアに投稿された緯度と経度に基づいて生成されたセル番号を単語とみなし，投稿時刻順に並べ替えたものを 1 文として扱う．この文を学習することにより，あるセルから次のセルへの移動をベクトルとして表現する．

ここで，図 2 の 3 通りの移動パターンを示す．本研究では，任意のセル α とセル γ は移動先の地域が類似しているときにコサイン類似度が高く出力される必要がある．図 2(a) に示すように，セル α とセル γ がそれぞれ共通のあるセル β への移動を行っているとき，セル α とセル γ は類似するセルとされる必要がある．この移動パターンのとき，従来の Skip-gram モデルと改良した Skip-gram モデルは，どちらもセル α とセル γ からセル β への移動を学習し，コサイン類似度を高く出力すると考えられるので，この移動パターンを学習することは本研究の目的において適切である．しかし，図 2(b) のように，セル α の移動先のセルとセル γ の移動前のセルが共通していたとき，または図 2(c) のように，セル α とセル γ の移動前のセルが共通していたとき，従来の Skip-gram モデルは移動前のセルも学習してしまうため，セル α とセル γ のコサイン類似度を高く出力すると考えられる．本研究では移動先のセルが類似するセルを発見することを目的としているため，この移動パターンを学習することは適切ではない．

よって，本研究では，改良した Skip-gram モデルによって図 2(a) の移動パターンに該当する，移動先の地域が類似しているセル α とセル γ を発見できることを示す．

本論文の構成は以下の通りである．2 章では，本研究の関連研究について述べる．3 章では，ユーザの移動を学習するための Skip-gram モデルについて述べる．4 章では，提案手法を用いたユーザの移動の分析と可視化について述べる．5 章では，提案手法と他の手法による出力結果の比較による評価を行う．6 章では，本論文のまとめと今後の課題を述べる．

2 関 連 研 究

2.1 Word2vec の応用

Word2vec のアルゴリズムを応用することで単語だけでなく，様々なデータをベクトルで表現する研究が盛んに行われている．

Shoji ら [7] は，ある特定の場所周辺のツイートを用いて，Word2vec の Skip-gram モデルのアルゴリズムのように，ある特定の場所を中心の単語，周囲のツイートに現れる単語を周辺の単語として，特定の場所の雰囲気についてベクトルで表現する手法を提案した．Dhingra ら [1] は，Twitter のデータを用いて，単語単位ではなく文字単位でツイートを学習することで，ツイートをベクトル化する手法を提案した．文字単位で学習するため，ソーシャルメディアの投稿の表記ゆれや，略語などの処理に適している．この学習によって作成されたベクトルから，ハッシュタグを予測した．Grover ら [2] は，ノードの近傍の特徴についてサンプリングする際に，ネガティブサンプリングの手法として，ランダムウォークを用いたグラフをベクトル化する手法を提案した．Madjiheurem ら [4] は音楽のコードを特徴ベクトルに変換する，Chord2Vec を提案した．Ristoski ら [6] は RDF グラフのエンティティの分散表現を作成する RDF2Vec を提案した．この手法では，グラフを一連のエンティティシーケンスに変換し，これを Word2vec のアルゴリズムを用いて，エンティティの分散表現を作成した．Vosoughi ら [8] はツイートの文字レベルで CNN-LSTM エンコーダデコーダを用いて，ツイートの学習を行った．吉田ら [10] は Word2vec を拡張した文書の特徴ベクトル学習手法である，Paragraph Vector モデルを用いて，観光スポットの特徴ベクトルを作成し，ある観光スポットにおいての，ユーザの主観的特徴と類似する観光スポットを検索する手法を提案した．

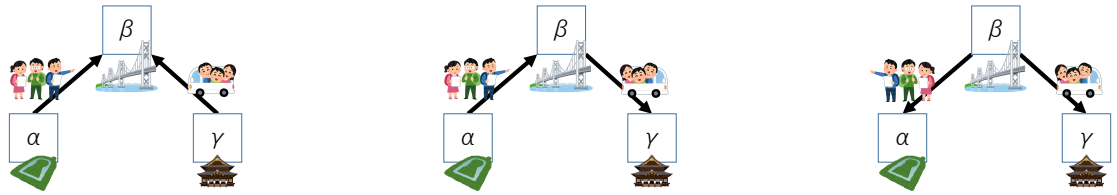
本研究で提案する手法も Word2vec の応用の一種である．本研究では，ツイートに付与された緯度経度情報を単語に変換することで，ある地域から次の地域への移動の関係について Word2vec を用いて学習し，ユーザの移動に基づく分散表現を作成する．

2.2 軌跡の応用

ソーシャルメディアサイトへの投稿からユーザの移動軌跡を抽出し，その分析を行う研究が行われている．

青山ら [11] はソーシャルメディアサイトに投稿された写真のジオタグと撮影日時を用いて，寄り道候補を発見する手法を提案した．この手法では，2 点間の移動の際に，人々の興味や知名度を考慮して寄り道の候補となる地点を発見した．谷ら [13] は Twitter のジオタグ付きツイートを用いて公共交通機関の交通路を抽出する手法を提案した．この手法ではユーザのツイートの投稿時間と緯度経度情報から移動速度の閾値を満たすツイートを抽出することで，何らかの交通手段を利用していると考えられる移動軌跡の近似直線をグループ化し，公共交通機関の交通路を抽出した．

本研究では，ソーシャルメディアサイトの投稿を用いて，2



(a) α と γ の移動先が共通　**(b)** α の移動先と γ の移動前が共通　**(c)** α と γ の移動前が共通

図 2: 移動パターンの例

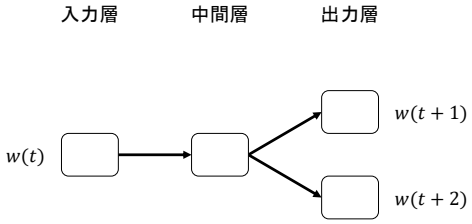


図 3: 改良した Skip-gram モデル

箇所の観光名所とその移動経路上に存在したユーザが次に移動した地域と移動先が類似する地点を発見することで，寄り道の候補となる地域の発見を行う．

3 提 案 手 法

本研究では，Word2vec の Skip-gram モデルを改良し，ユーザの移動軌跡からセル間のユーザの移動を学習し，セルのベクトルを作成する．

3.1 学 習 方 法

はじめに，従来の Skip-gram モデルについて説明する．図 1 に示す Skip-gram モデルでは，1 章で述べたように単語 $w(t)$ の周辺の単語を同じ文脈に出現する単語として学習を行う．本研究では，ユーザの移動軌跡について，あるセルを単語とみなして，あるセルから次のセルの移動のみを学習する．そのため，Skip-gram モデルをそのまま適用した場合，あるセル $w(t)$ について，$w(t)$ の移動前のセルである $w(t-c),\ldots,w(t-1)$ と，移動先のセルである $w(t+1),\ldots,w(t+c)$ の両方を学習することになるため，従来の Skip-gram モデルをそのまま適用することはできない．

次に，本研究で用いる改良した Skip-gram モデルについて説明する．本研究では，あるセルから，次のセルへの移動のみを学習するように改良した Skip-gram モデルを用いる．このモデルを図 3 に示す．この改良された Skip-gram モデルでは，ユーザの移動軌跡のあるセル $w(t)$ について，次の移動したセル $w(t+1),\ldots,w(t+c)$ を学習する．これにより，あるセル $w(t)$ のベクトルを作成する．このモデルは，Skip-gram モデル

表 1: ダミーデータの学習に用いたパラメータ

パラメータ	値
次元数	100
最小単語出現回数	10
ウィンドウサイズ	1
ネガティブサンプリング	5
学習反復回数	100
学習率	0.025

表 2: Skip-gram によるセル番号 “87” のコサイン類似度

順位	セル番号	類似度
1	“42”	0.308
2	“89”	0.279
3	“66”	0.277
4	“55”	0.252
5	“75”	0.250

表 3: 改良した Skip-gram によるセル番号 “87” のコサイン類似度

順位	セル番号	類似度
1	“0”	0.799
2	“79”	0.744
3	“75”	0.736
4	“12”	0.729
5	“100”	0.726

と同様のパラメータの設定が可能であるが，本論文では，あるセル $w(t)$ について，学習の対象とするのは直後の移動先のセルのみとする．そのため，セル $w(t+1)$ のみを学習する必要があるので，ウィンドウサイズは 1 とする．

従来の Skip-gram モデルでは，学習の際に文書中の単語の順番をランダムに入れ替える事によって，学習に用いる単語と同じ文書中のウィンドウサイズに含まれていない単語についても周辺の単語として学習を行い，単語のベクトルを作成している．しかし，本研究では一連の移動軌跡の中のある地域について，次の移動先の地域以外を学習することは，ある地域から直後の移動先の地域を学習してベクトルを作成するという点で適切ではない．そのため，学習の際に移動軌跡の順番を入れ替えないようにした．

3.2 ダミーデータによる実験

ここで，従来の Skip-gram モデルと，本研究の改良した Skip-

表 4: 図 2 のパターンごとの α と γ に対応する β

移動パターン	α	γ	β
図 2(a)	“87”	“42”	-
	“87”	“0”	“23”
	“87”	“0”	“57”
	“87”	“0”	“95”
図 2(b)	“87”	“42”	“64”
	“87”	“42”	“80”
	“87”	“0”	“95”
図 2(c)	“87”	“42”	“0”
	“87”	“42”	“48”
	“87”	“0”	“63”
	“87”	“0”	“64”

gram モデルの学習結果の差異を示すためにダミーデータを用いた実験を行う．この実験において，従来の Skip-gram モデルでは，あるセルの移動前のセルと移動先のセルを学習しているのに対して，改良した Skip-gram モデルでは，あるセルから次の移動先のセルのみを学習し，セルのベクトルを作成していることをセルごとのコサイン類似度によって確認することを目的とする．

本実験のダミーデータは，実際の移動軌跡のようなデータを表すため，“0” から “100” までのセル番号を用いてデータを作成した．これらのセル番号は 101 個の地域があることを表す．用意したセル番号の集合から，2 から 10 個をランダムに復元抽出したものを 200 件作成する．このとき，2 つのセルが連続して抽出された場合，これは移動していないことを表すので，連続したセル番号は 1 つのセル番号として扱った．また，それぞれのリストについてランダムで 1 倍から 50 倍に複製した．この作成したダミーデータは 5,431 件のリストとなった．このデータを用いて従来の Skip-gram モデルと，本研究の改良した Skip-gram モデルで学習させる．

ダミーデータについて，Skip-gram モデルと本研究の改良した Skip-gram モデルを用いて学習を行った結果について述べる．学習のパラメータには表 1 の値を用いた．それぞれの Skip-gram モデルで学習を行った結果，得られたセル番号のベクトルについて，コサイン類似度でそれぞれのセル番号間の類似度を求めた．本実験では任意に選択したセル番号 “87” についてのコサイン類似度の結果について述べる．表 2 に従来の Skip-gram モデルによる類似度の演算結果の上位 5 件，表 3 に改良した Skip-gram モデルによる類似度の演算結果の上位 5 件を示す．

従来の Skip-gram モデルによるコサイン類似度の演算結果が 1 位だったセル番号 “42” の移動先のセル番号と，改良した Skip-gram モデルによるコサイン類似度の演算結果が 1 位だったセル番号 “0” の移動先のセル番号がセル番号 “87” の移動先と共通するセル番号を図 2 のそれぞれパターンについて表 4 に示す．

まず，従来の Skip-gram モデルのセル番号 α =“87” とコサイン類似度が 1 位であったセル番号 γ =“42” の結果について述べる．移動先が同じ地点であることを表す図 2(a) の移動パターンでは共通する移動先の β に該当するセル番号は存在しなかった．また，本実験ではセル番号の学習の対象としない，図 2(b) と図 2(c) の移動パターンについては，セル番号 α =“87” とセル番号 γ =“42” で β に該当するセル番号が存在し，これらのセル番号について学習が行われたため，類似するセル番号として出力されたと考えられる．

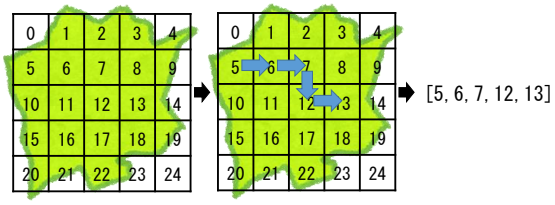


図 4: 移動軌跡のセル番号への変換

次に改良した Skip-gram モデルのセル番号 α =“87” とコサイン類似度が 1 位であったセル番号 γ =“0” の結果について述べる．移動先が同じ地点であることを表す図 2(a) の移動パターンでは共通する移動先の β に該当するセル番号が存在し，これらについて学習を行っている．セル番号 γ =“0” の移動先の件数のうち，β に該当するセル番号への移動件数の割合は，0.392 であった．また，図 2(b) と図 2(c) の移動パターンについては，β に該当するセル番号も存在しているが，Skip-gram モデルの改良を行った事により，学習は行われていない．

以上のことにより，従来の Skip-gram モデルは図 2(a) の移動パターンで β に該当するセル番号が存在しなかったため，移動先が類似するセル番号を出力することが不可能であったが，移動先のみを学習するように改良した Skip-gram モデルでは移動先が類似するセル番号 α とセル番号 γ をコサイン類似度を算出することによって出力が可能であることを示せた．

3.3 移動軌跡を用いた Skip-gram の学習

Skip-gram モデルを用いて単語のベクトルを作成するためには，学習を行う単語列が必要となる．そして，本研究の改良した Skip-gram モデルを用いてユーザの移動を学習するためには，あるユーザの一連の移動軌跡を 1 つの単語列として入力する必要がある．そこで，本研究では，図 4 のように分析する範囲をグリッドに分割し，ユーザの移動軌跡の緯度経度情報をグリッドに当てはめ，ユーザの移動軌跡をセル番号に変換することで，学習に用いるデータを作成する．

ユーザの投稿からセル番号に変換するためのデータを作成するための手順について述べる．はじめに，ユーザの一連の投稿について投稿時間順にソートする．次に，ある投稿から b 時間以内の投稿を 1 つの移動軌跡とする．本論文では，b は 1 時間とする．

次に，移動軌跡をセル番号に変換する方法について述べる．ユーザの移動軌跡の 1 点の緯度と経度をそれぞれ lat,

表 5: 実験に用いたパラメータ

パラメータ	値
次元数	100
最小単語出現回数	10
ウィンドウサイズ	1
ネガティブサンプリング	5
学習反復回数	50
学習率	0.25

lng とする．また，分析する範囲の緯度の最大値と最小値を lat_{max}, lat_{min}，経度の最大値と最小値を lng_{max}, lng_{min}，セルの分割数の縦，横をそれぞれ g_{lat}, g_{lng} とする．これらを用いて以下の式により，セル番号 yx に変換する．

$$yx = \frac{(lat - lat_{min}) \times g_{lat}}{lat_{max} - lat_{min}} \times g_{lng} + \frac{(lng - lng_{min}) \times g_{lng}}{lng_{max} - lng_{min}} \quad (1)$$

ユーザの移動軌跡から式 1 によって算出されたセル番号 yx を改良した Skip-gram モデルに入力する 1 文とする．

4 可 視 化

4.1 データセット

本実験に用いるデータセットについて説明する．Twitter から，2015 年 5 月 21 日から 2018 年 12 月 31 日の期間に投稿された，緯度経度情報が付与されているツイートを Twitter Streaming API[4]を用いて収集した．この中から，本実験では京都市周辺で投稿されたツイートを用いる．その際，これらのツイートから，ツイート数が 1 件のユーザ，緯度と経度が小数点第 3 位以下を持たないツイートは除外した．

また，Bot による自動投稿の影響を少なくするために，森國ら [12] の方法を参考に，前処理を行った．次の条件に当てはまるツイートを Bot による投稿として除いた．

- Twitter クライアント名に “NightFoxDuo” を含む．
- ツイート内容に “きつねかわいい !!! ” を含む．
- 緯度経度が (34.967096, 135.772691) である．
- Twitter クライアント名または表示名の 1 つ以上に “BOT”，“Bot”，“bot”，“人工無能” のいずれかを含む．

これらのツイートを除いた結果，ユーザ数は 107,944 人，ツイート数は 1,202,825 件であった．これらのツイートのデータを用いて実験を行った．

本実験では，セルの 1 辺を約 100m の四角形として作成した．その結果，縦 148 マス，横 189 マスの 27,972 セルが得られた．ここで，各セルごとにユーザ数が 2 人以下のセルは除外し，各移動軌跡中の連続する同じセル番号は 1 つのセル番号として扱う．

4.2 改良した Skip-gram によるモデル作成

本実験では，改良した Skip-gram モデルを用いてベクトルを作成した．学習のパラメータは，表 5 の値を用いた．パラメータの最小単語出現回数は，本研究では，各セルごとのツイート数であり，10 件以上のツイートがあるセルを対象として学習している．最終的に，学習に用いた移動軌跡数は 132,832 件でセル数は 3,045 セルとなった．

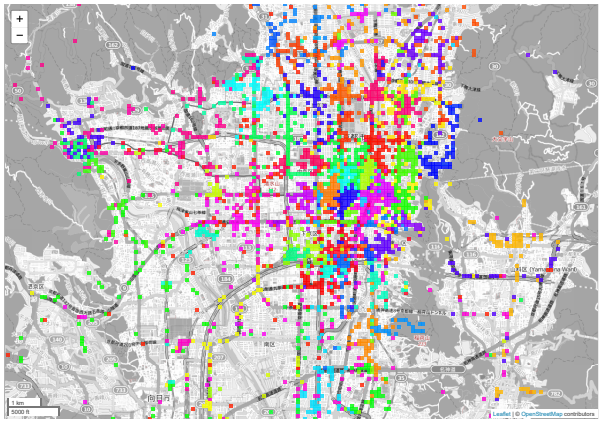


図 5: 京都市周辺のセルのクラスタリング結果

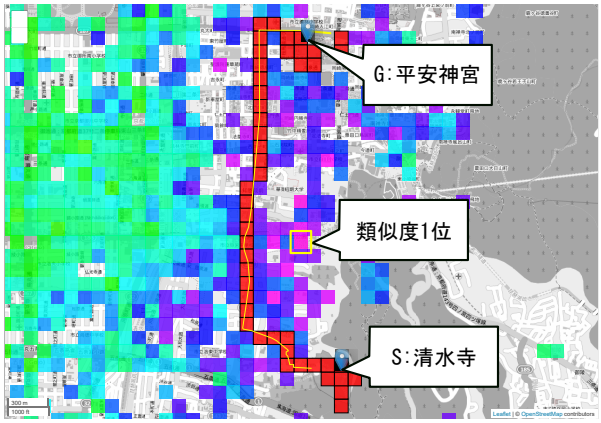


図 6: 清水寺から平安神宮までの移動経路と演算結果

4.3 可視化結果

改良した Skip-gram モデルで，データセットを学習し，作成されたベクトルを K-Means を用いてクラスタリングした結果を図 5 で示す．その際にクラスタ数は 100 とした．図 5 において 1 つのクラスタを 1 色で表している．地図は OpenStreetMap[5] を用いた．これらのクラスタで同じ色で示されているセルは，ベクトル間のコサイン類似度が高いため，次の移動先の地域が類似している地域を表していると考えられる．

また，このベクトルに対しての意味演算を行った．本実験では，観光名所を 2 箇所選択し，その観光名所と移動経路上のセルを抽出する．それらの全てのセルの加算を行い，そのベクトルについて，コサイン類似度の高いセルを出力した．この演算により，2 箇所の観光名所とその移動経路上のユーザが次に移動する地域と，次の移動先の地域が類似する別の地域を発見する．

4 : https://dev.twitter.com/streaming/overview

5 : https://www.openstreetmap.org/

表 6: Precision, NDCG による評価実験結果

	Precision			NDCG		
	@1	@3	@5	@1	@3	@5
改良した Skip-gram	0.400	0.489	0.501	0.400	0.501	0.513
Skip-gram	0.400	0.428	0.430	0.400	0.429	0.432
移動確率	0.500	0.450	0.417	0.500	0.455	0.429

京都市の観光名所を TripAdvisor[6]を基に選択し，2 つの観光名所間の移動経路は Openrouteservice[7]から取得した．本実験では，2 箇所の観光名所として，清水寺と平安神宮を選択した．移動経路として取得した情報は，自動車による移動としている．清水寺周辺と平安神宮周辺とその移動経路を赤色，また，それらのセルを全て加算した結果の類似度を可視化した結果を図 6 に示す．

意味演算の結果，類似度の高いセルは清水寺や平安神宮の付近，その間の移動経路付近のセルに多く出力された．コサイン類似度が 1 位と算出されたセルについて，清水寺や平安神宮の付近，その間の移動経路上のセルとのコサイン類似度は，0.85 であった．このセルの投稿内容を分析するために，範囲に含まれるツイートについて，MeCab[8]を用いて形態素解析を行った．この際に，辞書は mecab-ipadic-NEologd[9]を用いて普通名詞と固有名詞を抽出した．このセルには “坂本龍馬” や “円山公園” といった形態素が出現し，清水寺や平安神宮などに興味を持つユーザが観光をする際に，寄り道の候補となることが考えられる．

選択した観光名所や，その移動経路付近のセルがコサイン類似度が高く算出された理由として，本研究では，ユーザのツイート投稿の間隔が 1 時間以内であるものを一連の移動軌跡としていることがあげられる．また，改良した Skip-gram モデルによってある地域から次の移動先の地域のみを学習しているため，コサイン類似度が高く算出されたセルは，次の移動先が移動経路上のセルと類似している．そのため，この演算の結果，出力されたセルは，演算に用いた観光名所や移動経路の周辺の地域であり，その地域に滞在したユーザと同様の興味を持つユーザが存在した地域である．したがって，類似度が高く出力されたセルの地域は，観光名所間の移動の際に立ち寄ることができる地域であると考えられる．

5 評価実験

本章では，2 箇所の観光名所とその移動経路上のセルを入力することによって，寄り道の候補となる地域を発見する事において，提案した手法が有効であることを示す．

5.1 実験条件

本実験では，提案手法である改良した Skip-gram モデルと従来の Skip-gram モデルと入力されたセルの移動確率を用いた手法の 3 つの手法によって実験結果の比較を行った．

移動確率を用いた手法について説明する．学習に用いたセル番号の集合を G，集合 G に含まれる各セル番号を g，入力された移動軌跡上のセル番号の集合を R ($R \in G$)，移動軌跡中の各セル番号を r として，セル番号 g における移動確率の合計 $S(g)$ を以下の式によって求める．

$$S(g) = \sum_{r \in R} \sum_{g \in G} P(r, g) \tag{2}$$

ここで，$P(r, g)$ は移動軌跡中のセル r からセル g への移動確率を表す．この手法の結果は，移動確率の合計 $S(g)$ の値が高い順に出力する．

評価実験には，4.1 節で使用したものと同じデータを使用した．

2 つの観光名所の選定方法は，京都市の観光名所を TripAdvisor の “京都市の名所・見どころ” によるランキングの上位 10 件に含まれる観光名所をランダムに選択した．その観光名所間の移動の 10 ルートを用いて，それぞれの手法に観光名所付近とルート上のセルを入力し，出力されるセルの上位 5 件について，そのセルに含まれる地域が 2 点間の移動中に寄り道に適切かどうかを人手で判断した．寄り道のできる地域の判断の条件として，移動経路の付近に存在していること，またそのセルの範囲に寄り道に適切であると思われるスポットが存在していることとした．

5.2 評価指標

本実験では，評価指標として，Precision@k，Normalized Discounted Cumulative Gain Measure@k (NDCG@k) [3] の 2 つを用いた．

Precision@k は，検索結果の上位 k 件において，評価の高いものが含まれているほど，ランキングの評価は高くなる．本論文の場合は，上位 k 件に，寄り道としてふさわしいと評価されたセルの数であり，以下の式で算出する．

$$Precision@k = \frac{1}{k} \sum_{i=1}^{k} rel_i \tag{3}$$

ここで，rel_i は i 番目の順位の評価点数を表す．

NDCG@k は，ランキングの上位 k 件において，理想的なランキングへの近さを表す評価指標である．Precision@k との違いとして，NDCG@k は，ランキングの順位による重みが付与されている．ランキングの上位 k 件において，同じ件数だけ良い評価のものが存在した場合は，上位に出現するほどランキングは高く評価される．NDCG を計算するための Discounted Cumulative Gain Measure (DCG) は以下の式で算出される．

$$DCG@k = rel_1 + \sum_{i=2}^{k} \frac{rel_i}{\log_2 i} \tag{4}$$

DCG@k は順位の下降とともに重みを付け加えている．NDCG@k は，以下の式で算出される．

6 : https://www.tripadvisor.jp/
7 : https://openrouteservice.org/
8 : http://taku910.github.io/mecab/
9 : https://github.com/neologd/mecab-ipadic-neologd

$$NDCG@k = \frac{DCG@k}{IDCG@k} \quad (5)$$

ここで IDCG@k は，DCG が最も理想的だった場合の値を表す．本実験では，$k = \{1, 3, 5\}$ とする．

5.3 評価結果

はじめに，$k = 1$ の評価実験の結果，Precision，NDCG による評価結果は，移動確率を用いた手法の結果が最も良い結果を示している．このため，1 位のみの出力では，移動確率を用いて寄り道の候補を発見することが有効であるという結果になった．また，改良した Skip-gram モデルと従来の Skip-gram モデルの 1 位として出力されたセルの評価値は 10 ルート中 8 ルートで同じ評価値であった．

次に，$k = \{3, 5\}$ の評価実験の結果，Precision，NDCG による評価結果は改良した Skip-gram モデルが最も良い結果を示した．$k = 1$ のときに有効であった移動確率を用いた手法は，出力結果に駅を含むセルが多く含まれており，観光名所間の移動の際に通る地点であるが，寄り道の候補として適切ではないと考えられる．

この結果より提案した手法が上位の複数のセルを寄り道の候補として出力することにより，寄り道の候補となる地域の発見において有効であることを示した．

6 おわりに

本論文では，ユーザの 2 点間の移動について分析するために，ユーザの移動に基づく分散表現を作成する手法を提案した．提案手法では，ユーザの緯度経度情報をセルで表現し，あるセルから次の移動先のセルのみを学習するように改良した Skip-gram を用いて学習を行った．京都市周辺における Twitter のデータを利用して学習を行い，観光地間の移動の最中に寄り道の候補となる地域を可視化した．また，本研究の提案手法と従来の Skip-gram モデルと入力されたセルの移動確率を用いた手法の 3 つの手法を用いて評価実験を行った．評価の結果，本研究の提案手法が複数の寄り道の候補を出力した際に有効であることを示した．

今後の課題として，提案手法を用いた観光地間の移動の分析についてより多くのパターンにおいて実験を行い，考察を行うことがあげられる．また，本研究では，ある地域を表すセルから移動先のセルへの 2 点間の移動を用いて移動先の類似するセルの発見を行ったが，n 点間の移動を 1 単語として扱うことで，類似する移動のパターンを発見することがあげられる．

謝辞

本研究は，首都大学東京傾斜的研究 (全学分) 学長裁量枠戦略的研究プロジェクト戦略的研究支援枠「ソーシャルビッグデータの分析・応用のための学術基盤の研究」及び JSPS 科研費 16K00157, 16K16158 による．

文献

[1] Bhuwan Dhingra, Zhong Zhou, Dylan Fitzpatrick, Michael Muehl, and William W Cohen. Tweet2vec: Character-based distributed representations for social media. In *The 54th Annual Meeting of the Association for Computational Linguistics*, 2016.

[2] Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 855–864. ACM, 2016.

[3] Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems*, Vol. 20, No. 4, pp. 422–446, 2002.

[4] Sephora Madjiheurem, Lizhen Qu, and Christian Walder. Chord2vec: Learning musical chord embeddings. In *Proceedings of the Constructive Machine Learning Workshop at 30th Conference on Neural Information Processing Systems*, 2016.

[5] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pp. 3111–3119, 2013.

[6] Petar Ristoski and Heiko Paulheim. Rdf2vec: Rdf graph embeddings for data mining. In *International Semantic Web Conference*, pp. 498–514. Springer, 2016.

[7] Yoshiyuki Shoji, Katsurou Takahashi, Martin J Dürst, Yusuke Yamamoto, and Hiroaki Ohshima. Location2vec: Generating distributed representation of location by using geo-tagged microblog posts. In *International Conference on Social Informatics*, pp. 261–270. Springer, 2018.

[8] Soroush Vosoughi, Prashanth Vijayaraghavan, and Deb Roy. Tweet2vec: Learning tweet embeddings using character-level cnn-lstm encoder-decoder. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '16, pp. 1041–1044. ACM, 2016.

[9] Z. Yu, H. Xu, Z. Yang, and B. Guo. Personalized travel package with multi-point-of-interest recommendation based on crowdsourced user footprints. *IEEE Transactions on Human-Machine Systems*, Vol. 46, No. 1, pp. 151–158, 2016.

[10] 吉田朋史, 北山大輔, 中島伸介, 角谷和俊. ユーザレビュー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, 2017.

[11] 青山賢, 廣田雅春, 石川博, 横山昌平. ジオタグ付き写真を用いた知名度が低いにもかかわらず興味の度合いが高い寄り道候補の発見. 第 7 回データ工学と情報マネジメントに関するフォーラム, 2015.

[12] 森國泰平, 吉田光男, 岡部正幸, 梅村恭司. ツイート投稿位置推定のための単語フィルタリング手法. 情報処理学会論文誌データベース（TOD）, Vol. 8, No. 4, pp. 16–26, 2015.

[13] 谷直樹, 風間一洋, 榊剛史, 吉田光男. ジオタグ付きツイートを用いた交通路の抽出. 第 7 回データ工学と情報マネジメントに関するフォーラム, 2015.

[14] 野津直樹. ビッグデータによる観光動態分析 (特集 観光情報学). 人工知能：人工知能学会誌：Journal of the Japanese Society for Artificial Intelligence, Vol. 31, No. 6, pp. 850–857, 2016.