群衆の移動履歴とカテゴリを用いた特徴抽出に基づく地理オブジェクト検索システム

大塚　公貴†　　北山　大輔††　　角谷　和俊†

† 関西学院大学総合政策学部　〒669-1337 兵庫県三田市学園 2 丁目 1
†† 工学院大学情報学部システム数理学科　〒163-8667 東京都新宿区西新宿 1-24-2
E-mail: †{ezn23068,sumiya}@kwansei.ac.jp, ††kitayama@cc.kogakuin.ac.jp

あらまし　位置情報サービスなどの普及により, ユーザが地理情報検索を行うことは容易になりつつある. しかし, 既存の検索システムは地理オブジェクトの利用目的に基づいた検索を行うには適していない. 例えば,「食べ歩き中の観光に適したお寺」や「買い物中の観光に適したお寺」など, 地理オブジェクトは同じカテゴリに属していても利用目的が異なる場合がある. このような利用目的から, 既存の検索システムで検索を行う事は困難である. そこで本研究では, ユーザの行動履歴とカテゴリを用いて地理オブジェクトの持つ利用目的を推定する. ユーザは我々の提案するシステムを通して, 利用目的に基づいた検索を行うことが可能となる.

キーワード　地理情報, 情報検索, 観光情報, 人の移動と行動, 情報推薦

1. はじめに

ユーザが観光情報や地理情報を取得する為に, 旅行ブログや SNS などの情報リソースを利用する場合がある. 観光情報サイトじゃらん (注1)ではホテルの予約や観光地のレビューを閲覧することが可能である. また, Instagram(注2)では, 位置情報やハッシュタグを用いた検索が可能であるため, 様々な地理オブジェクトに対して情報の取得が可能である. このようなサービスの登場により, 地理オブジェクトの情報取得は容易になりつつあるが, 既存の検索システムは地理オブジェクトの利用目的に基づいた検索には適していない.

地理オブジェクトは同じカテゴリに属していても利用目的が異なる場合がある. 例えば, 三重県の伊勢神宮は周囲におかげ横丁などの食べ歩きに利用されるオブジェクトが多数存在することから「食べ歩き中の観光に適したお寺」と捉えることができる. 一方で, 神戸市に位置する湊川神社は周囲に元町商店街があることから「買い物中の観光に適したお寺」と捉えることができる. このように同じ「寺・神社」というカテゴリに属する地理オブジェクト同士であってもその利用目的は異なる場合がある. このように, ユーザより詳細な利用目的を基に検索を行う場合, 既存の検索システムで検索を行うのは困難である. そこで本研究では, ユーザの行動履歴とカテゴリを用いて地理オブジェクトの持つ利用目的を推定する. ユーザは我々の提案するシステムを通して, 利用目的に基づいた地理オブジェクト検索を行うことが可能となる.

2. 関連研究

本章では，関連する地理情報検索の研究や群衆の行動履歴を用いた研究について紹介し, 本研究の位置付けを明らかにする.

2.1 地理情報に関する研究

地理情報を用いて情報推薦を行う研究は盛んに行われている. 石野らは, ブログデータベースから旅行ブログエントリを検出し, その中から観光情報として, 土産物情報と観光名所情報を抽出する手法を提案した [1]. 旅行ブログを分析対象としている点で参考にしている. また, 土田らは Word2Vec を用いた意味的演算を観光分野に用いることにより, 都市・地域とランドマークの意味的な関係性を捉える研究を行った [2]. 地理オブジェクトに対する意味演算は利用目的の推定を行う際の参考としているが, 土田らは観光分野に焦点を当て分析を行っているのに対し, 我々は観光地以外のオブジェクトに対しても焦点を当て分析を行っているため, 本研究とは趣旨が異なる. また, Li らはロケーションの曖昧さに注目し, ニュース記事と旅行ブログの共起分析に基づく地理情報検索を支援するシステムを提案した [3]. 本研究とは地理情報の検索支援を行うという点で類似しているが, Li らがロケーションの曖昧さに着目しているのに対し, 我々はオブジェクトの利用目的に着目しているため, 本研究の趣旨とは異なる.

2.2 地理情報検索に関する研究

地理情報検索の精度向上を図る研究も盛んに行われている. 手塚らは WWW 上のテキストデータに対する内容解析によって人間の地理空間認知の構造を明らかにし, 地域情報検索の効率化を行った [4]. また, 廣嶋らは地理情報検索のクエリ入力を支援する仕組みとして指定した場所の特徴的な語の提示を行い，場所の指定だけでキーワードを想起させる手法を提案した [5]. 他にも McCurley らは Web 上の地理的コンテキストを発見し, コンテキストに対する内容理解の支援を行う手法を提案した [6]. これらの研究は地理情報検索の精度向上という点に

(注1)：https://www.jalan.net
(注2)：https://www.instagram.com

おいて, 参考にしている.

2.3 ユーザの移動履歴を用いた研究

地理情報の研究の中にはユーザの行動履歴を用いた物も存在する. 山本らは, web 上に投稿された位置情報を用いて, 時間と空間を関連付け, ユーザの現在地を考慮したリアルタイムのナビゲーションを行うシステムの構築を行った [7]. また, 篠田らは GPS などの位置探知デバイスから自動的に得られるユーザの移動情報を利用し, 移動の特徴を分析することでユーザ間の類似度を算出し, 協調フィルタリング手法に適用するという手法を提案した [8]. これらの研究はユーザの移動情報を用いるという点で参考にしているが, 地理情報検索の精度向上を図る本研究とは趣旨が異なる. 他にも Li らユーザの履歴に基づいてユーザー間の類似性を地理的にマイニングする研究を行った [9]. また, Cheng らは地域広告のターゲティングの為に, Twitter の位置情報サービスを利用したユーザの分析を行った [10]. これらの研究は移動データを分析したものであるが, 分析対象がユーザであり, 我々は分析対象を地理オブジェクトとしている点が異なる.

3. システム概要

本稿では, ユーザの地理オブジェクトに対する利用目的に基づく検索システムを提案する (図 1). まず入力としてユーザは検索を行いたい領域と訪問したいオブジェクトのカテゴリ, オブジェクトの利用目的の三つの要素を選択する. ここでは, ユーザが三宮周辺の名所・史跡でかつグルメ・レストランにも訪れられるスポットを検索したいという要求を持っていたと仮定する. 次にシステムは入力された利用目的を基に検索ベクトルを生成する. 最後に検索ベクトルとオブジェクトが持つ特徴ベクトルとのコサイン類似度を用いてユーザの要求するオブジェクトを出力する. ここでは, 三宮周辺の名所で飲食も行えるオブジェクトとして南京町や生田神社などが出力されている. 我々の提案するシステムにより, ユーザは同じエリアやカテゴリに属する地理オブジェクトであっても, 利用目的から地理オブジェクトを検索することが可能となる.

4. 提 案 手 法

「対象の地理オブジェクトがどのカテゴリと共に利用されるか」ということが, 対象オブジェクトの利用目的を表すという仮説のもと検索システムの構築を行う. 本研究では, 地理オブジェクトの利用目的の抽出を行うためにユーザの移動履歴とカテゴリに着目した. 例えば, ある地理オブジェクト A に対する群衆の移動データにカテゴリ X が頻出していたとする. この場合, 対象の地理オブジェクト A はカテゴリ X と共に利用される傾向が高く, ユーザが地理オブジェクト A に対して持つ利用目的もカテゴリ X と類似する可能性がある. 本研究では, 群衆の移動データと地理カテゴリの二つの要素を用いることで地理オブジェクトの利用目的を推定し, 検証を行う.

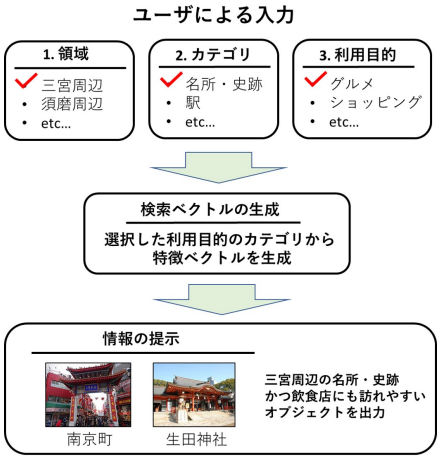

図 1 システム概要

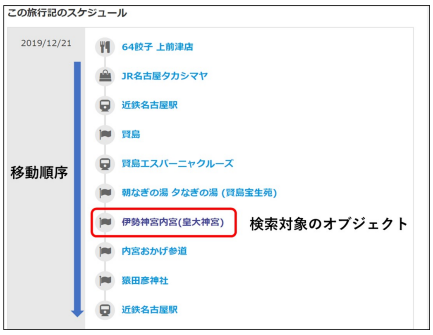

図 2 4travel の旅行ブログ

4.1 移動データの取得

群衆の移動データの取得には 4travel(注3)の旅行ブログを使用した. 旅行ブログとは旅行中の見聞や体験を書き記した文書のことである. 我々は旅行ブログの「地理オブジェクト間の移動を時系列で表す」という特性に注目し, これを移動データとして用いた. 図 2 はあるユーザの旅行ブログである. 旅行ブログを確認するとユーザの一日の行動履歴を取得することができる. 例えば, このユーザは 「グルメ・レストラン」 →「デパート」→ … →「寺・神社」→「寺・神社」→「寺・神社」→「駅」といったカテゴリ間の移動を行っている.

4.2 地理オブジェクトの特徴ベクトルの抽出手法

本節では地理オブジェクトの特徴ベクトルの抽出手法について述べる. 地理オブジェクトの特徴ベクトルの算出には前節で述べた, 4travel の移動データを用いた. 本研究では対象の地理オブジェクトと地理カテゴリの共起度を用いることにより, 地理オブジェクトの利用目的を推定する. 利用目的の推定には対象の地理オブジェクトに対するユーザの前後の移動を用いた Hop 関数を用いた (式 1). x は対象の地理オブジェクトを表しており y は移動先の地理オブジェクト,c はカテゴリを表す. また, 本研究では対象の地理オブジェクトに対する前後 4 つの移動を用いて特徴量の算出を行う. 従って hop_num は 1 から 4 の間を取る.

(注3) : https://4travel.jp

カテゴリ	オブジェクト	重み
駅	近鉄名古屋駅	1/4
自然・景勝地	賢島	1/3
乗り物	賢島エスパーニャクルーズ	1/2
温泉	朝なぎの湯	1/1
対象のオブジェクト	伊勢神宮	
名所・史跡	内宮おかげ参道	1/1
寺・神社	猿田彦神社	1/2
駅	近鉄名古屋駅	1/3
該当なし	該当なし	1/4

表 1　ユーザの移動に対するオブジェクトの重み

$$Hop(x) = \sum \frac{1}{hop\_num(y, c)} \tag{1}$$

表 1 は, 図 2 から取得した移動に式 1 を用いた出力結果である. 対象の地理オブジェクトの直近に訪れられるカテゴリほど地理オブジェクトの利用目的を表すという仮説のもと, ユーザの移動に対して重み付けを行った. 従って, 対象の地理オブジェクトからの移動順が離れていけば離れていくほど, カテゴリの重要度は減少していく.

本研究では, 一つの地理オブジェクトにつき 30 件の旅行ブログをユーザの移動として取得した. 表 2 はその出力結果を表す. 一列目は 4travel のカテゴリを表しており, 本研究ではカテゴリ数の 20 次元で地理オブジェクトの特徴ベクトルを表現した. 出力結果を見ると伊勢神宮はカテゴリ「名所・史跡」にあたるオブジェクトと共に利用される傾向が高く, 逆に「ショッピングモール」や「アウトレット」などとは共に利用されにくいということがわかる.

本節では, 地理オブジェクトがどのカテゴリと共に利用されるのかということを表現するための Hop 関数について述べた. 次節では, カテゴリ自体が持つデータの偏りとその重み付けについて述べる.

4.3　カテゴリの重み付け

本節ではカテゴリに対する重み付け手法について述べる. 前節で述べた地理オブジェクトが持つカテゴリの重み付け算出手法では, ユーザの利用傾向が高いカテゴリのスコアが高くなるという問題がある. 例えば, 図 2 は伊勢神宮の Hop 値の出力結果である. 出力結果を見るとカテゴリ「名所・史跡」のスコアが極端に高いことがわかる. 旅行ブログは旅先の見聞や体験を記したものであるため, 観光地や飲食店, 公共交通機関の出現頻度が高くなってしまう. そこで, 我々はカテゴリ自体に対する重み付けとして $TF-iDF$ における iDF と同様の手法を用いた. $TF-iDF$ とは, 文書中に含まれる単語の重要度を算出する手法の一つである. 本研究では iDF における全文書数 N をユーザの総数に, 文書の出現頻度を表す DF をユーザの出現頻度を表す $UF(UserFrequency)$ として定義し, カテゴリの重みを算出した (式 2).

$$iUF = \log\left(\frac{N}{UF}\right) \tag{2}$$

式 2 を用いて出力したものが図 3 である. 本研究では各カテゴリに属する地理オブジェクトの満足度順上位 5 件の地理オブジェクトに対してユーザの移動の取得を行った. また各地理オブジェクトに対して 30 件の旅行ブログを収集しユーザの移動を取得した. その結果, ユーザ数は 2412 件となり, UF(各カテゴリの出現頻度) は図 3 のようになった. データを見ると, 「グルメ・レストラン」や「名所・史跡」,「駅」などのカテゴリが頻出していることがわかる. これらの UF 値に対して式 2 を用いることにより, iUF(inverse User Frequency) 値の算出を行った. 結果を見ると, 出現頻度の低い「アウトレット」や「百貨店・デパート」の重みが高くなり, 出現頻度の高いカテゴリの重要度は低くなっていることがわかる.

カテゴリ	Hop 値
駅	8.00
温泉	0.58
テーマパーク	0.25
動物園	0.33
ビーチ	1.00
美術館・博物館	2.33
公園・植物園	0.50
名所・史跡	20.75
寺・神社	0.25
自然・景勝地	4.42
百貨店・デパート	0.33
スーパー・コンビニ・量販店	0.00
ショッピングモール	0.00
アウトレット	0.00
市場・商店街	0.00
お土産屋・直売所・特産	1.00
グルメ・レストラン	8.92
乗り物	3.58
道の駅	0.58
空港	0.00

表 2　伊勢神宮の Hop 値

4.4　オブジェクトの利用目的の推定

ここまで, カテゴリとの共起度を図る尺度である Hop 値と, カテゴリ自体に重みを付与する iUF 値という二つの関数の説明を行った. 本節ではこの二つの要素を用いた地理オブジェクトの利用目的推定手法について述べる.

$$Hop \times iUF = \sum \frac{1}{hop\_num(y, c)} \times \log\left(\frac{N}{UF}\right) \tag{3}$$

式 3 により, ユーザの行動履歴とカテゴリの重みに基づく地理オブジェクトの利用目的の推定を行った (図 4). Hop 値と $Hop \times\ iUF$ 値のスコアを見てみると「名所・史跡」などは変わらず数値が高いが, カテゴリとして, 出現頻度の高い「グルメ・レストラン」は数値が低くなっている. このスコアから伊勢神宮は寺・神社というカテゴリに属する地理オブジェクトであるが, 利用用途として, 観光に用いられたり, 風景を楽しむために訪れられる地理オブジェクトであると捉えることができる.

カテゴリ	UF 値	iUF 値
駅	824	0.47
温泉	144	1.22
テーマパーク	92	1.42
動物園	99	1.39
ビーチ	112	1.33
美術館・博物館	218	1.04
公園・植物園	301	0.90
名所・史跡	1108	0.34
寺・神社	349	0.84
自然・景勝地	363	0.82
百貨店・デパート	21	2.06
スーパー・コンビニ・量販店	50	1.68
ショッピングモール	148	1.21
アウトレット	10	2.38
市場・商店街	103	1.37
お土産屋・直売所・特産	63	1.58
グルメ・レストラン	1373	0.24
乗り物	344	0.85
道の駅	122	1.30
空港	392	0.79
N(取得ユーザ合計)	2412	

表 3　カテゴリの重要度

カテゴリ	Hop 値	iUF 値	$Hop \times iUF$ 値
駅	8.00	0.47	3.73
温泉	0.58	1.22	0.71
テーマパーク	0.25	1.42	0.35
動物園	0.33	1.39	0.46
ビーチ	1.00	1.33	1.33
美術館・博物館	2.33	1.04	2.44
公園・植物園	0.50	0.90	0.45
名所・史跡	20.75	0.34	7.01
寺・神社	0.25	0.84	0.21
自然・景勝地	4.42	0.82	3.63
百貨店・デパート	0.33	2.06	0.69
スーパー・コンビニ・量販店	0.00	1.68	0.00
ショッピングモール	0.00	1.21	0.00
アウトレット	0.00	2.38	0.00
市場・商店街	0.00	1.37	0.00
お土産屋・直売所・特産	1.00	1.58	1.58
グルメ・レストラン	8.92	0.24	2.18
乗り物	3.58	0.85	3.03
道の駅	0.58	1.30	0.76
空港	0.00	0.79	0.00

表 4　伊勢神宮の $Hop \times iUF$ 値

5. 評価実験

前章では地理オブジェクトの利用目的推定の為に, Hop 値と iUF 値を用いて地理オブジェクトの特徴ベクトルの算出例を示した. 本章では, 算出された $Hop \times iUF$ 値について評価を行う. 評価の方法として, 我々は略地図のラベルが「地理オブジェクト集合の利用目的を表現する」という特性に注目した.

図 3　仙台市の略地図

仙台観光マップ	大崎八幡宮	JR仙台駅	瑞鳳殿	資料展示館	仙台城跡	仙台博物館	宮城県美術館
大崎八幡宮		0.61	0.87	0.79	0.89	0.82	0.90
JR仙台駅	0.61		0.56	0.34	0.71	0.70	0.67
瑞鳳殿	0.87	0.56		0.82	0.81	0.91	0.91
青葉城資料展示館	0.79	0.34	0.82		0.63	0.73	0.75
仙台城跡	0.89	0.71	0.81	0.63		0.74	0.87
仙台博物館	0.82	0.70	0.91	0.73	0.74		0.91
宮城県美術館	0.90	0.67	0.91	0.75	0.87	0.91	

表 5　コサイン類似度 (仙台市の略地図)

図 3 は仙台市の略地図である.

略地図の作成者は略地図に対し「仙台観光マップ」というラベルを用いている. このことから, 略地図に出現している地理オブジェクト群は「観光」という利用目的で用いられると推測できる. 略地図中に出現している各地理オブジェクトの $Hop \times iUF$ 値を算出し, それぞれのコサイン類似度を算出する. 略地図中に出現している地理オブジェクト群は略地図のラベルに基づき出現している為, 地理オブジェクト間の類似度は全体的に高くなることが予想される. 仮に算出されたコサイン類似度が全体的に低い場合, $Hop \times iUF$ 値はユーザの利用目的を表す特徴量として正しく表現できていないと判断できる.

実験はエリアや縮尺の異なる略地図, 10 件を収集し分析を行った. 表 5 に分析結果の一例を記す. 表 5 から JR 仙台駅を除いた地理オブジェクト間で高い類似度が確認できる.

他の略地図でも同様にオブジェクトのコサイン類似度を算出した結果, 全体的に類似度は高くなった.

6. 考察と今後の課題

本章では評価実験の考察と今後の課題について述べる. コサイン類似度の算出結果を見ると, JR 仙台駅だけが極端に他のオブジェクトとの類似度が低い (図 5). これは駅というオブジェクトの特性上, 公共交通機関としてユーザが利用をしており, 地図のラベル「観光」には合致しないオブジェクトであるからだと推測できる. 一方で, 略地図中に出現する JR 仙台駅以外のオブジェクト間での類似度は, 比較的高くなっていることから, これらのオブジェクトは略地図のラベル「観光」に適したオブジェクト群であることが推測される. 以上の結果から $Hop \times iUF$ 値は, オブジェクトの利用目的に則した特徴量を付与することが可能であるとわかった.

本研究では 4travel の旅行ブログをユーザの行動履歴データとして扱い, オブジェクトの利用目的の推定に用いたが, 既存のデータセットには偏りがある. 本研究の目的は地理検索システ

ムの構築であるため, 今後は分析データを追加し, 観光分野以外での応用を目指す.

謝　　辞

本研究の一部は，令和元年度科研費基盤研究 (A)(課題番号：16H01722)，基盤研究 (B)(課題番号：19H04118) によるものです.

文　　献

[1] 石野亜耶, 難波英嗣, 竹澤寿幸, ” 旅行ブログエントリからの観光情報の自動抽出”, 知能と情報:日本知能と情報ファジィ学会誌,vol.22,No.6, pp.667-669(2010)

[2] 土田崇仁, 遠藤雅樹, 加藤大受, 江原遥, 廣田雅春, 横山昌平, 石川博, “Word2vec を用いた地域やランドマークの意味演算”, 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)，H5-1

[3] H. Li, R. K. Srihari, C. Niu, and W. Li. infoXtract location normalization: a hybrid approach to geographical references in information extraction. In Workshop on the Analysis of Geographic References, Edmonton, Canada, May 2003. NAACL-HLT.

[4] 手塚 太郎, 李 龍, 高倉弘喜, 上林弥彦, “Web の内容解析に基づく地理的領域の特性付けと地域情報検索への応用”, 情報処理学会研究報告データベースシステム, 2002, 67(2002-DBS-128), 503 - 508

[5] 廣嶋 伸章, 安田 宜仁, 藤田 尚樹, 片岡 良治, “地理情報検索におけるクエリ入力支援のための特徴語の提示”, 人工知能学会,2012, 1C1-R-5-6

[6] K. S. McCurley. Geospatial mapping and navigation of the web. In Proc. of the 10th int. conference on World Wide Web, pages 221–229. ACM Press, 2001.

[7] 山本 浩司, 安村 禎明, 片上 大輔, 新田 克己, 相場 亮, 宮城 政雄, 桑田 仁, “ユーザの投稿情報に基づく経路ナビゲーション”, 人工知能学会全国大会論文集, 2004, pp1-4

[8] 篠田 裕之, 竹内 亨, 寺西 裕一, 春本 要, 下條 真司, “行動履歴に基づく協調フィルタリングによる行動ナビゲーション手法”, 情報処理学会研究報告マルチメディア通信と分散処理（DPS), 2007, 91(2007-DPS-132), pp87 - 92

[9] Li, Q. and Zheng, Y. et al. Mining user similarity based on location history. In Proc. of GIS ’ 08 (Santa Ana, CA, Nov. 2008). ACM Press: 298-307

[10] Zhiyuan Cheng, James Caverlee, and Kyumin Lee. A content-based approach to geo-locating twitter users. In Proceedings of the 19th ACM International Conference on Information and Knowledge Management, pp. 759–768. ACM, 2010