

観光の形態を考慮した旅行者の将来の訪問国予測

柴田 有基¹, 石野 亜耶², 難波 英嗣³, 竹澤 寿幸¹

¹ 広島市立大学大学院

² 広島経済大学

³ 中央大学

¹ {shibata, takezawa}@ls.info.hiroshima-cu.ac.jp

² ay-ishino@hue.ac.jp

³ nanba@kc.chuo-u.ac.jp

概要: 本研究では、ソーシャルメディアから収集した旅行者の過去の訪問国履歴データを用い、旅行者の将来の訪問国を予測する手法を提案する。旅行者の過去の訪問国から将来の訪問国を正確に予測するためには、過去の訪問国の文化や位置情報のほかに、観光の特徴などを考慮することで旅行者の好みを把握する必要がある。そこで、本研究では、深層学習手法の LSTM を用い、Wikipedia から獲得した国の情報に加え、観光の形態という観点からソーシャルメディアを解析して得られる特徴を考慮することで、より正確な将来の訪問国の予測を試みる。

Keywords: 時系列データ, 深層学習, 観光地推薦

1. はじめに

現在、COVID-19 の影響により、海外旅行をはじめとする観光行動が減少している。しかし、これまでの訪日外国人の増加傾向や行動自粛へのストレスなどを考慮すると、今後、海外旅行者数が回復していくことが期待される。そのため、海外旅行を考えている人へのマーケティング活動は非常に重要になると考えられる。

観光分野では、ソーシャルメディアを解析し、観光スポットや観光ルートなどを推薦する研究は既に行われている。しかし、多くの研究は、ある都市やある地域などの比較的小さな範囲を対象としているため、世界中の国が対象となる海外旅行者への対応は不十分である。そこで、本研究では、世界中の国を対象とし、旅行者の過去の訪問国履歴から将来の訪問国を予測する。これにより、世界中を対象とした観光地推薦が可能になるほか、予測結果からライバルとなりうる国がわかることでお互いの国の観光政策に活かすことが可能になる。

本論文の構成は以下の通りである。まず、2 章では関連研究について述べる。次に、3 章では、観光の形態を考慮した旅行者の将来の訪問国の予測について述べる。また、4 章では、3 章で紹介した手法による実験について述べる。そして、最後に 5 章

で本論文のまとめを述べる。

2. 関連研究

Su ら[1]は、ユーザの訪問履歴に加えて、フレンド関係や家が近いユーザの訪問履歴を考慮した予測手法を提案している。また、Baraglia ら[2]は、ユーザの特徴と PoI (Point of Interest) の特徴を考慮した機械学習手法の Ranking SVM (Support Vector Machine) によって将来訪れる PoI を予測する手法を提案している。これらの研究に対して、本研究では、深層学習手法を用いる点で異なる。

深層学習を用いた研究では、Liu ら[3]は位置情報と時間情報を考慮した、次の訪問地の予測手法を提案している。また、Kong ら[4]は、Liu らの手法を拡張し、観光客の訪問地の周期性を考慮した手法を提案している。これらの研究では、ある国や地域内のデータを対象としているが、本研究では世界中の国を対象としている点で異なる。

国の推薦に関する研究では、Majoddi ら[5]は、移民のための移住国推薦システムを提案している。また、Johnson ら[6]は、旅行者向けに国ごとの腸チフス感染症の危険性を考慮した、予防接種の推薦システムを提案している。これらの研究に対し、本研究では、観光地の観光情報を研究対象としている点で異なる。

表 1: 観光の形態の定義と具体例 (柴田ら[9]より引用)

観光の形態	定義	例
インフラ, ハードツーリズム[10]	近代的な建造物や娯楽施設を対象にした観光.	橋, ダム, テーマパーク, ショッピングモール, 水族館, 博物館, 動物園
ヘルスツーリズム[11]	心身を癒すことや散歩などの軽い運動を通して健康維持を目的とした観光.	宗教的巡礼, 温泉, ハイキング, トレッキング
スポーツツーリズム[12]	スポーツを体験または観戦することを目的とした観光.	オリンピック, 野球・サッカー観戦
グリーンツーリズム[13]	自然と触れ合うことを目的とした観光.	農業 (漁業) 体験, フルーツ狩り, ピクニック
ヘリテージツーリズム[13]	世界遺産や歴史的な建築物を対象にした観光.	世界遺産, 国宝, 寺, 神社, 城
カルチュラルツーリズム[14]	それぞれの地域の生活や文化, 民族, 伝統などを対象にした観光.	着物体験, 神楽, 祭り, 初詣

3. 観光の形態を考慮した旅行者の将来の訪問国の予測

本研究では, ソーシャルメディアから構築した訪問国履歴データセットを用い, 旅行者の過去の訪問国から将来の訪問国を予測する. より正確に将来の訪問国予測を行うためには, 過去の訪問国から旅行者の特徴や好みを獲得し, 考慮する必要がある. そこで, 本研究では, 深層学習手法の LSTM (Long short-term memory) [7]を用い, それぞれの国の Wikipedia から得られる情報および観光の形態の特徴を考慮した予測を行う.

3.1. Wikipedia を用いた国の情報の獲得

旅行者が訪問国を決定する際には, 国の位置情報や文化, 公用語などを考慮することが考えられる. そこで, 本研究では, このような国の情報を Wikipedia から獲得し, 将来の訪問国の予測に用いる. Wikipedia の情報の獲得には, Wikipedia2Vec[8]の事前学習モデルを利用する. Wikipedia2Vec とは, Wikipedia のデータを用いて, 単語の分散表現とエンティティの分散表現を同一空間で学習・表現する手法であり, 手軽に Wikipedia の情報を獲得することができる. 本研究では, この Wikipedia2Vec の事前学習モデルから, それぞれの国名や地域名に該当するエンティティの分散表現を獲得し, これを LSTM の入力データに用いる.

ところで, 国の位置情報を考慮する場合には, アメリカにはハワイ州やアラスカ州, グアム島などが含まれるように, 地理的に離れている地域が存在する国がある. このような場合, 予測に位置情報を考慮することが難しくなるが, 本研究ではこの問題に対し, 離れている地域を分割することで対応する.

3.2. 観光の形態の特徴

近年の観光では, 旅行者は多様な目的を持っていることが多いため, その訪問先で体験できる観光の特徴は, 訪問国の決定の際の重要な判断材料になると考えられる. そこで, 本研究では, 観光の特徴の一つとして, 柴田ら[9]で提案されている表 1 の観光の形態を利用し, 国ごとに投稿されたソーシャルメディアを解析することで, 得られた結果を国の観光の特徴として予測に用いる.

柴田らの研究では, ソーシャルメディアである旅行ブログエントリをそれぞれの観光の形態に分類することで, 解析を行っている. 本研究も同様に, 国や地域ごとにソーシャルメディアを解析し, その結果を考慮した予測を行う. 具体的には, 解析に用いたソーシャルメディアの内, 観光の形態に分類された割合を各観光の形態で算出し, この 6 次元の分散表現をその国の観光の特徴とする.

4. 評価実験

4.1. 実験条件

【実験に用いるデータ】

TravelBlog (<https://www.travelblog.org/>) の 2002 年 4 月 5 日から 2019 年 1 月 18 日までに投稿された約 500,000 件の旅行ブログエントリを実験に用いる. また, TravelBlog では, 著者自身がどの地域についての記事であるかを設定する場合があります. 設定すると国名や地域名が URL に反映される. 本研究では, この URL から旅行者の訪問国を抽出し, データセットを構築する. なお, アカウントによっては短期間で同一国についての投稿を複数存在する場所が見られたことから, 90 日以内に同じ訪問国が存在する場合はまとめる処理を行った. また, 各アカウントの訪問国の数がある程度必要と

なることから、6 개국以上の訪問国履歴があるアカウントのみを用いる。実験に用いるデータセットの詳細を表 2 に示す。本実験では、9,437 個のアカウントに対して、80%を訓練データに、残りの 20%を検証・評価データに用いる。また、アカウントごとの訪問国数の平均は 12.9 となった。

表 2: 実験に用いるデータセットの詳細

訓練データのアカウント数	7,549	9,437
検証データのアカウント数	944	
評価データのアカウント数	944	
アカウントごとの訪問国数の平均	12.9	

【機械学習】

表 2 のデータセットを用い、学習・検証・評価を行う。LSTM の主なパラメータについては、最適化アルゴリズムを Adam, バッチサイズを 2,048, 学習率を 0.001, LSTM 層を 1 層, 出力ユニット数を入力ユニット数と同じ数とし、検証データによる実験から epoch 数を決定する。

Wikipedia2Vec の事前学習モデルには、公開されているモデルで最も軽量な 100 次元のモデル (http://wikipedia2vec.s3.amazonaws.com/models/en/2018-04-20/enwiki_20180420_nolg_100d.pkl.bz2) を用いる。これは、次元数の異なるいくつかのモデルで予備実験を行ったところ、実験結果への影響がほとんど確認できなかったためである。また、観光の形態への分類については、機械学習手法の SVM と LightGBM[15]を用い、国ごとに収集した最大 800 件のブログエントリを各観光の形態に分類する。なお、機械学習の入力データには、ブログエントリのテキストに含まれる名詞のみとする。テキストデータの前処理については、fastText[16]の言語判定モデルを用い、英語と判定されたブログエントリのみを分類対象とした。また、品詞タグ付けに RNNTagger[17]を用いることで、名詞のみに絞る処理を行った。また、二つの機械学習手法については、観光の形態ごとに精度が高い手法を採用した。

【実験手法】

ベースライン手法と提案手法に対して、評価指

標 $\text{recall}@k$ ($k=1, 5, 10, 20$) を算出し、比較を行う。それぞれの手法については以下の通りである。

ベースライン手法

- Random: 無作為に選んだ国を予測結果とする。
- StarSpace: 汎用ニューラルモデルである StarSpace[18]を用い、協調フィルタリングの問題として、予測モデルを構築する。
- N-gram ($N=2$): 訓練データから bi-gram の予測モデルを構築する。
- Wikipedia2Vec: 過去の訪問国を、それぞれ Wikipedia2Vec の事前学習モデルを用いて分散表現に変換し、これらの分散表現の平均と cosine 類似度が高い国を予測結果とする。
- Distance: 最後に訪れた国から距離が近い国を予測結果とする。
- Most Popular: 訓練データで最も頻出した国を予測結果とする。
- LSTM(Wiki): Wikipedia2Vec の事前学習モデルを用いて、国や地域の分散表現を獲得し、これを入力データとした LSTM により予測を行う。

提案手法

- LSTM(Wiki+Type): Wikipedia2Vec の事前学習モデルから獲得した分散表現に、6 つの観光の形態の割合を加えた 106 次元の分散表現を入力データとした LSTM により予測を行う。

4.2. 実験結果と考察

評価実験の結果を表 3 に示す。これより、ベースライン手法の LSTM(Wiki) と提案手法の LSTM(Wiki+Type) で高い評価値が得られた。このことから、過去の訪問国履歴から将来の訪問国を予測する際には、LSTM が有効であることがわかる。また、LSTM 以外では、ベースライン手法の“Wikipedia2Vec”で比較的良い値が得られた。これにより、旅行者が訪問国を決定する際に、国の位置情報だけでなく、国の文化や公用語などの情報も考慮していることが考えられる。

ここから、LSTM(Wiki) と LSTM(Wiki+Type) の結果から、明確な有意差が確認できなかったことについて考察する。この原因の一つについては、Wikipedia2Vec と観光の形態の割合を結合した分

表 3: 将来の訪問国予測の評価実験の結果

	手法	recall@1	recall@5	recall@10	recall@20
ベースライン手法	Random	0.8%	2.3%	4.1%	8.3%
	StarSpace	0.9%	9.9%	19.8%	32.5%
	N-gram ($N=2$)	5.1%	-	-	-
	Most Popular	6.9%	22.5%	34.6%	51.8%
	Distance	7.1%	20.1%	26.9%	35.2%
	Wikipedia2Vec	9.8%	24.6%	35.9%	48.9%
	LSTM(Wiki)	<u>23.0%</u>	<u>53.6%</u>	<u>68.5%</u>	<u>81.1%</u>
提案手法	LSTM(Wiki+Type)	22.9%	<u>53.7%</u>	<u>68.5%</u>	<u>81.1%</u>

散表現から、二つの入力データの関係性を LSTM が捉えきれていないことが考えられる。この問題に対して、それぞれの入力データごとに処理を行い、その後の結合の方法を工夫することである程度改善することが考えられる。また、解析対象のブログエントリが国ごとに最大 800 件としたことから、データ数の不足も原因の一つに考えられる。

5. まとめ

本研究では、旅行者の過去の訪問国履歴を用い、Wikipedia の情報および観光の形態を考慮した、将来の訪問国の予測手法を提案した。旅行ブログエントリを用いた評価実験では、評価指標の $\text{recall}@k$ ($k = 1, 5, 10, 20$) において、観光の形態を考慮することの有効性は確認できなかったが、LSTM が有効であることが確認された。

今後は、LSTM のそれぞれの入力データの学習方法を工夫するなど、引き続き将来の訪問国の予測に観光の形態が有効であるかを検証していく。また、旅行者情報の追加などを行い、より精度の高い予測を目指していく。

参考文献

- [1] Y. Su, X. Li, W. Tang, J. Xiang, and Y. He, “Next Check-in Location Prediction via Footprints and Friendship on Location-Based Social Networks,” Proc. of 19th International Conference on Mobile Data Management, MDM’18, pp.251-256 (2018)
- [2] R. Baraglia, C. I. Muntean, F. M. Nardini, and F. Silvestri, “LearNext: learning to predict tourists movements,” Proc. of 22nd ACM international conference on Information & Knowledge Management, CIKM’13, pp.751-756 (2013)
- [3] Q. Liu, S. Wu, L. Wang, and T. Tan, “Predicting the next location: A recurrent model with spatial and temporal context,” Proc. of 30th AAAI Conference on Artificial Intelligence, AAAI-16, pp.194-200 (2016)
- [4] D. Kong and F. Wu, “HST-LSTM: A Hierarchical Spatial-Temporal Long-Short Term Memory Network for Location Prediction,” Proc. of 27th International Joint Conference on Artificial Intelligence, IJCAI-18, pp.2341-2347 (2018)
- [5] A. E. Majjodi, M. Elahi, N. E. Ioini, and C. Trattner, “Towards Generating Personalized Country Recommendation,” Proc. of 28th ACM Conference on User Modeling, Adaptation and Personalization, ACM UMAP 2020, pp.71-76 (2020)
- [6] K. J. Johnson, N. M. Gallagher, E. D. Mintz, A. E. Newton, G. W. Brunette, and P. E. Kozarsky, “From the CDC: new country-specific recommendations for pre-travel typhoid vaccination,” Journal of Travel Medicine, Vol.18, pp.430-433 (2011)
- [7] S. Hochreiter and J. Schmidhuber, “LONG SHORT-TERM MEMORY,” Journal of Neural computation, Vol.9, No.8, pp.1735-1780, (1997)
- [8] I. Yamada, A. Asai, J. Sakuma, H. Shindo, H. Takeda, Y. Takefuji, and Y. Matsumoto, “Wikipedia2Vec: An Efficient Toolkit for Learning and Visualizing the Embeddings of Words and Entities from Wikipedia,” arXiv:1812.06280 [cs.CL] (2018)
- [9] 柴田有基, 篠田広人, 難波英嗣, 石野亜耶, 竹澤寿幸: 観光の形態に基づいた旅行ブログエントリの自動分類と可視化, 観光と情報, Vol.16, No.1, pp.49-61 (2020)
- [10] 藤井千賀子, 茂木直美, 林由利子, 柳沼しほ: インフラツーリズムガイド 2018, 芸文社, (2018)
- [11] 河行茜, 木下藤寿: 島根おおだ健康ビューローの取り組み, 生涯スポーツ実践研究年報: 鹿屋体育大学生涯スポーツ実践センター所報, Vol.17, pp.28-35 (2019)
- [12] 高橋義雄, 原田宗彦, 岡星竜美, 工藤康宏, 二宮浩彰, 松岡宏高, 山下玲, 青木淑浩: スポーツツーリズム・ハンドブック, 学芸出版社 (2015)
- [13] 山下晋司: 観光学キーワード, 有斐閣 (2011)
- [14] 後藤和子: 観光と地域経済 -文化観光の経済分析を中心に-, 地域経済学研究, Vol.34, pp.41-47 (2018)
- [15] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, “LightGBM: A Highly Efficient Gradient Boosting Decision Tree,” Advances in Neural Information Processing Systems 30, NIPS 2017, pp. 3149-3157 (2017)
- [16] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, “Bag of Tricks for Efficient Text Classification,” arXiv:1607.01759 [cs.CL] (2016)
- [17] H. Schmid, “Deep Learning-Based Morphological Taggers and Lemmatizers for Annotating Historical Texts,” Proc. of 3rd International Conference on Digital Access to Textual Cultural Heritage, DATeCH 2019, pp.133-137 (2019)
- [18] L. Wu, A. Fisch, S. Chopra, K. Adams, A. Bordes, and J. Weston, “StarSpace: Embed All The Things!,” arXiv:1709.03856 [cs.CL] (2017)