地域特有の観光情報の推薦手法の提案

労 瑛瑩　魏 逸倫　韓 東力
日本大学

1 はじめに

国土交通省観光庁[1]の統計により，2017 年に日本人国内旅行の消費額は 21 兆 1,028 億円となり，国内延べ旅行者数も 6 億 4,720 人に達した。その中で，個人旅行者としての観光客数はかなりの割合を占めていると思われる。また,「観光戦略実行推進タスクフォース」の開催や，地方への観光魅力発見や体験型観光への関心集めなど様々な対策を検討するなど，政府が積極的に個人旅行の促進を実施している[2]。個人旅行の旅行者が旅行代理店を通さず，自らがネットにおいて取得できた観光情報をもとに旅行プランを策定するのは一般的である。しかし、この方法ではかなりの時間とコストがかかってしまうため，旅行者のニーズに当てはまる観光情報を効率的に提供することが求められている。

そこで，ウェブ上に蓄積されている観光情報を抽出し，旅行者のニーズを満たす情報提供を試みる研究が盛んに行われている。中でも特に旅行ブログに着目し，文書に出現した語句に対する解析をもとに観光情報を抽出･分類するものが多かった。例えば，佐藤らは観光客に類似体験型観光スポットの情報を提供することを目指し，ある体験スポットに対し類似度の高い地域を自動的に検出するツールを構築した[3]。徳久らは，観光地の特徴を共起グラフで簡素化して表示する手法を提案した[4]。また，遠藤らは地域観光サイトとブログの観光情報を抽出・融合した上で提示する手法を提案した[5]。

しかし，いずれの既存研究においても、抽出された観光情報と地域との関連度を細分化していないという問題がある。地域に関わる観光情報は 2 種類あり，この地域にしか提供されない観光情報、および他の地域にも現れる観光情報である。そこで，地域の観光魅力の最大化に利用されるアピールポイントとして、地域に特有の観光情報を検出することが重要と思われる。我々は旅行ブログを研究対象として，地域の観光情報を抽出・分析することで，ユーザーに地域特有の観光情報を提示するシステムを開発することを目指している。本稿では，このような枠組みにおける最初の一環として、独自の方法で旅行ブログの集合から地域特有の特徴語を選定し，それを利用することでその地域に特有の観光情報を検出・階層的に推薦する手法を提案する。

2 使用するデータ

本稿では，旅行記コーパスとしてフォートラベル（http://4travel.jp/）から 2004 年 4 月 1 日～2018 年 10 月 21 日に投稿された国内旅行記（ブログ）における七つの地域のブログデータ(合計 196,649 件･各地域コーパスのブログ件数は表 1 に示す）を使用する。単語の分散表現は Wikipedia をもとに Word2Vec[6]を用いて構築した。

本研究では，東京を代表例として地域コーパスから地域特有の特徴語抽出を行い，東京における特色のある観光情報を検出し，ユーザーに提供する。

表 1 ：地域ごとのブログ件数

地域	ブログの件数
東京	55,508
広島	9,607
北海道	28,805
京都	30,734
名古屋	19,669
沖縄	29,329
静岡	22,997

提案手法の有効性が確認されれば，東京以外の地域にも簡単に拡張できると思われる。

3 提案手法

地域において特色のある観光情報を推薦する手順は以下 4 つのステップからなる。

Step1.　地域特有の特徴語の選定
Step2.　地域特有性スコアの算出
Step3.　旅行ブログの観光内容の分類
Step4.　地域特有の観光情報の推薦

以下では、それぞれのステップについては詳細を述べる。

Copyright(C) 2019 The Association for Natural Language Processing.
All Rights Reserved.

3.1 地域特有の特徴語の選定

地域特有の特徴語選定手法として，本研究では単語（名詞のみ）の文書出現頻度 DF（Document Frequency）に着目する。ある単語wの地域コーパスにおける文書出現頻度とコーパス全体における文書出現頻度の比$ratio_w$がこの単語が当該地域の特徴語としての特有性の強さを表す。

$$\mathrm{DF}_{w,all} = \mathrm{DF}_{w,tokyo} + \mathrm{DF}_{w,\sim tokyo} \quad (1)$$

$$ratio_w = \frac{\mathrm{DF}_{w,tokyo}}{\mathrm{DF}_{w,all}} \quad (0 < ratio_w \leq 1) \quad (2)$$

式(1)と(2)により，単語wの$\mathrm{DF}_{w,tokyo}$が$\mathrm{DF}_{w,\sim tokyo}$を大きく上回れば，$w$が東京に出現する確率は高いことが分かり，東京の特有な観光情報を表す言葉であると推定できる。従って、$ratio_w$＞0.5 を満たす単語を東京地域に特有の特徴語として選定する。

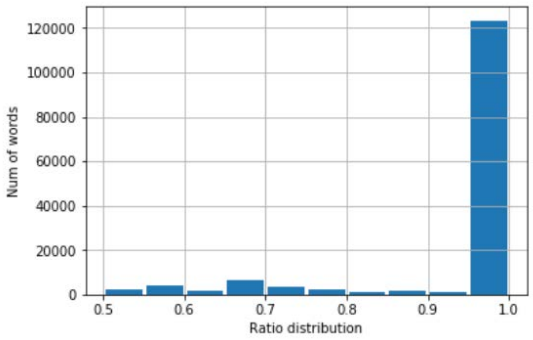

図 1 ：東京地域に特有の特徴語の分布

図 1 には東京の地域コーパスから選定された地域特有の特徴語の分布を示している。特徴語の総数は 147,561 個であり，X 軸は$ratio_w$を表しており（刻みは 0.05 である），Y 軸は各区分内に含まれる単語数を表している。

表 2 ：2 種類の単語群の例

強関連単語群
日比谷公園大音楽堂，昭和女子大，吉祥寺バウスシアター，浅草食通街，高尾山口，上野東照宮，奥多摩湖
弱関連単語群
大江戸，日枝神社，海老蔵，徳川家光，海蔵寺，聖天宮，水月観音

図 1 の分布結果をもとに，特徴語を地域との関連度の強さにより 2 種類の単語群に分ける。強関連単語群は$ratio_w = 1$である単語の集合であり，東京地域コーパスにしか出現しないので，地域の特有性が最も強いといえる。一方，東京地域以外のコーパスにも現れているため，弱関連単語群は$0.5 < ratio_w < 1$を満たす単語の集合である。表 2 に強関連単語群と弱関連単語群の具体例が挙げられている。3.2 節で，この 2 種類の単語群を用い，ブログに含まれる地域特有の観光情報量を計測する手法を述べる。

3.2 地域特有性スコアの算出

3.1 節で得られた結果に基づき，東京の旅行ブログごとに地域の特色ある観光情報量を表すスコア（以下では地域特有性スコアと呼ぶ）を算出する。

$$score_d = \sum_{w \in W_d} \frac{freq_w}{N_d} \times ratio_w \quad (3)$$

式（3）では，N_dとW_dがそれぞれブログ d に含まれる総単語数と単語の集合を表し，$freq_w$は単語wがブログ d における出現頻度を表す。

図 2 は地域特有性スコアの分布図である。X 軸は東京の地域特有性スコアの分布範囲であり，Y 軸はスコア区間（0.1 刻み）ごとに含まれるブログ件数を表す。特有性スコアが大きければ，当該ブログに含まれる地域特有の特徴語が多くなり，すなわち，記載されている観光内容は当該地域に限られた観光情報である可能性が高いと考えられる。したがって，特有性スコアの分布を利用すれば，地域に特有な観光情報を階層的に提供することが可能である。

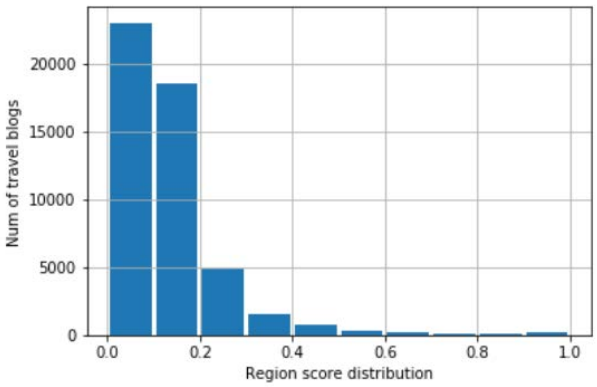

図 2 ：地域特有性スコアの分布

3.3 観光内容の分類

本節では，ブログに含まれる観光内容を分類する手順を述べる。文書解析にはトピック分析を用いる

Copyright(C) 2019 The Association for Natural Language Processing.
All Rights Reserved.

のが一般的な手法である。既存研究では，旅行ブログに対する分類におけ LDA の利用が有効であることが示されている[7]。この考え方に基づき，本研究では LDA を用いて東京の旅行ブログに対しトピック分析を行い、予め選定しておいた観光内容体系のカテゴリに得られたトピックを割り当てることで、ブログに書かれた観光内容を分類する。具体的には，まず単語の分散表現により，各トピックの上位 100 個のトピックワード集合における中心語と各カテゴリの上位 100 個の類似語集合における中心語をそれぞれ算出する。そして，各トピックの中心語と各カテゴリの中心語の距離を計算することにより，カテゴリとトピックの対応関係を求める。

ここで，観光サイトじゃらん net（https://www.jalan.net/）と既存研究[4][5]により，3 種類の観光内容体系が候補として選定され，表 3 にそれぞれの内容を示す。LDA ではトピック数 K=10，50，100 と 3 種類設定したため，合計 9 種類の組み合わせが得られた。

表 3 ：3 種類の観光内容体系

体系 1	交通,スポーツ,エンタメ・アミューズメント，レジャー・体験，自然・景観・絶景，風呂，ショッピング，観光・施設，宿泊，料理
体系 2	移動，携行，見聞，飲食，体験，購入，交流，宿泊
体系 3	見る・遊ぶ，祭り・イベント，自然・文化，食べる・泊まる，土産・特産

本研究では，地域と関連度の強い観光内容を検出することを重視しているため，3.2 節で説明した地域特有性スコアとカテゴリに属す観光内容との関連性の高さを同時に考慮することで，最も妥当と思われるトピックとカテゴリの組み合わせを目指す。上述した 9 種類の組み合わせのそれぞれにおいて，各カテゴリに対しサンプルブログをランダムに抽出し，人手でカテゴリ分類の正確さを分析した。

図 3 は Topic 数 K＝50, 観光内容体系 1 で「料理」カテゴリに属すブログの観光内容関連度の散布図である。X 軸はブログがカテゴリに属す度合いを表し，Y 軸は地域特有性スコアを示す。図 3 は 4 つの区域に分けることができ，各区域から 1 つだけサンプルブログを抽出する。4 つの区域にそれぞれ「カテゴリとの関連性が低い･地域特有性スコアが低い」,「カテゴリとの関連性が低い･地域特有性スコアが高い」,「カテゴリとの関連性が高い・地域特有性スコアが低い」と「カテゴリとの関連性が高い・地域特有性スコアが高い」ブログが含まれていると考えられる。各区域から抽出された 4 つのサンプルブログに対し、人手でブログの内容をチェックし，属している区域の特徴を満たすかどうかを判別する。区域の特徴が満たされれば正解と判定し，これらの手順を繰り返し，観光内容体系における全てのカテゴリについて検証することで，平均正解率を算出する。

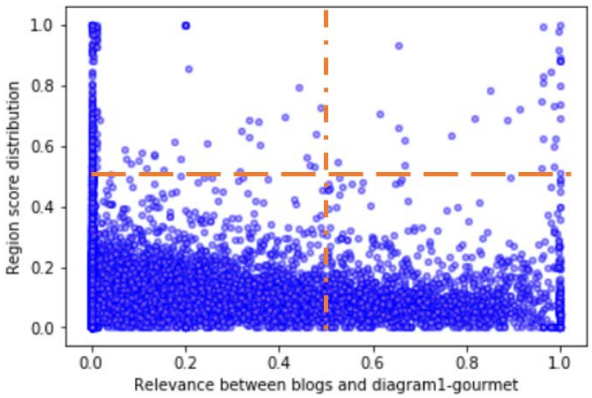

図 3 ：ブログの観光内容関連度の分布

検証を行なった結果，トピック数 K＝50 と観光内容体系 3 の組み合わせで平均正解率が最も高かったため、後続の処理では体系 3 を用いる。表 4 は，観光内容体系 3 のカテゴリと各カテゴリに割り当てられるトピック数を示している。

表 4 ：各カテゴリに割り当てされたトピック数

観光内容体系 3 のカテゴリ	トピック数 K=50
見る・遊ぶ	10
祭り・イベント	10
自然・文化	12
食べる・泊まる	10
お土産・特産品	8

3.4 地域特有の観光内容の推薦

3.2 節及び 3.3 節で得られた結果を用いて、地域観光情報を地域特有性スコアをもとに段階的に推薦していく。具体的な推薦事例を 4 章で述べる。

Copyright(C) 2019 The Association for Natural Language Processing.
All Rights Reserved.

4 ブログの推薦事例

本章では，東京特有の観光情報を含むブログの推薦事例を一つ挙げる。4.1 節では，ブログの推薦基準と評価方法，4.2 節では，推薦結果に対する考察を述べる。

4.1 ブログの推薦基準と評価方法

ここで，東京の地域コーパスから，テストデータとして合計 14,923 件の旅行ブログを抽出し，これらを対象に提案手法を用いて推薦事例の抽出を行う。

ブログの推薦条件は，指定されたカテゴリとの関連性が高く，かつ高い地域特有性スコアを持っていることである。今回はカテゴリとの関連性の値が 0.9 に達し，さらに地域特有性スコアが 0.9 以上あることを推薦条件とする。推薦されたブログに対する妥当性の評価は、著者の一人により地域特有の観光情報が含まれているか否かを手動で判定することで行われる。

4.2 推薦結果の考察

表 5 に，各観光カテゴリにおいて推薦基準を満たすブログの抽出結果と人手による判定結果を示す。

表 5 ：推薦事例

カテゴリ 3	推薦されたブログ数	判定結果	
		○	×
見る・遊ぶ	16	14	2
祭り・イベント	15	12	3
自然・文化	15	13	2
食べる・泊まる	4	3	1
お土産・特産品	4	4	0

「×」と判定されたブログに、観光カテゴリとの関連性がないものがほとんどである。一方,「○」と判定されたブログには東京と関連性の高い固有地名やイベント名などが書かれている。例えば,「自然・文化」カテゴリで推薦されたブログは「国営昭和記念公園」,「皇居東御苑」,「神代植物公園」といった東京固有の地名を含んでいる。また,「見る・遊ぶ」カテゴリにおいては、東京タワーやスカイツリーなどの場所名を含むブログが推薦された。

5 おわりに

本研究では，旅行ブログから地域特有の特徴語の選定について新たな手法を提案した。取得した特徴語に基づき，ブログごとに地域特有性スコアを付けた上で，地域の特色ある観光情報を含むブログの抽出を行った。推薦されたブログを手動で分析した結果より，提案手法を用いて地域特有の観光情報を階層的に検出できることが示唆された。今後はブログ推薦の評価実験を行い，本提案手法の有用性を検証していく。また，本手法に基づき観光情報を階層的に推薦するシステムの構築を目指す。

参考文献

[1] 国土交通省観光庁ホームページ http://www.mlit.go.jp/kankocho/

[2] 首相官邸ホームページ，政策会議，明日の日本を支える観光ビジョン構想会議 https://www.kantei.go.jp/jp/singi/kanko_vision/index.html

[3] 佐藤菜摘，難波英嗣，石野亜耶，竹澤寿幸：類似観光スポットの比較による魅力発見システムの構築，観光情報学会 第 14 回研究発表会, Vol.12, 2016 年.

[4] 徳久雅人，竹中直人，木村周平，谷本圭志：トップダウン型共起グラフを用いたブログからの観光地の行動分析，第 23 回言語処理学会年次大会発表論文集(Web)，Vol.23rd，p. 20-24，2017 年.

[5] 遠藤雅樹，中村信也，奥秋清次，大野成義：地域サイト及びブログからの観光情報抽出と融合の提案，情報処理学会研究報告，Vol.2012-DBS-155, No.6，p.1-6,2012 年.

[6] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In Proceedings of Workshop at ICLR, 2013.

[7] 宋紫龍，古屋秀樹：訪日中国旅行者の旅行記を用いた旅行情報抽出方法の基礎の分析，第 32 回日本観光研究学会, 全体大会学術論文集, p.109-112, 2017 年.

Copyright(C) 2019 The Association for Natural Language Processing.
All Rights Reserved.