ISS-SP-057

観光内容の地域関連度の細分化手法

労　瑛瑩††　　魏　逸倫†　　韓　東力†
† 日本大学情報科学科　　†† 日本大学大学院総合基礎科学研究科

1. はじめに

近年，ウェブ上に蓄積されているブログから，観光情報を抽出・提供する研究が盛んに行われている[1][2]。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が細分化されていないという問題がある。本稿では，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ブログに含まれる観光内容の地域関連度を細分化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日までに投稿された国内旅行記(ブログ)における「東京，広島，北海道，京都，名古屋，沖縄，静岡」七つの地域のブログデータ(合計196,649件)を使用する。

3. 提案手法

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を抽出し，ブログに含まれる観光内容と東京との地域関連度を数値化する。処理手順は以下の3つからなる。

3.1 地域特有の特徴語の選定　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。$ratio_w > 0.5$ を満たす単語を東京地域に特有の特徴語として選定し，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。表1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。

表1:2種類の単語群の例

強関連単語群	弱関連単語群
日比谷公園大音楽堂	大江戸
浅草食通街	日枝神社
高尾山口	徳川家光
上野東照宮	聖天宮
奥多摩湖	水月観音

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)を式(3)より算出する。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。

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

図1は東京コーパスにおける地域特有性スコアの分布図である。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。

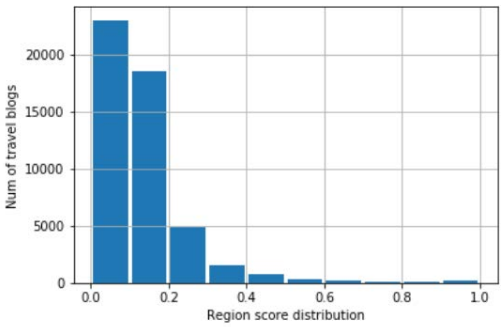

図1. 地域特有性スコアの分布

3.3 観光内容の分類　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[2]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。その後，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，ブログ観光内容の地域関連度を細分化していく。

5. まとめ

本研究では，旅行ブログから提案手法で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与した上で，ブログ観光内容の地域関連度の細分化を試みた。今後は本提案手法の有用性を検証していくと共に，本手法に基づき地域特有性の観光情報を段階的に推薦するシステムの構築を目指す。

参考文献

[1] 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. 20-23，2017 年.

[2] 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.