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
Classification and Visualization of Travel Blog Entries Based on Types of Tourism

柴田有基
Naoki Shibata

広島市立大学大学院 情報科学研究科
Graduate School of Information Sciences, Hiroshima City University

概要 本研究では，旅行ブログエントリ中のテキスト，画像および Wikification の結果を用いて 6 種類の観光の形態に分類する手法を提案する．また，これらの情報を，深層学習ベースの手法で統合する手法を提案し，ensemble 学習を用いた実験では，精度 0.807 を得た．最後に，観光の形態から地図上にマッピングされた旅行ブログエントリを検索できるシステムを構築した．

1　はじめに

近年，観光は従来の娯楽を追求するのみだけではなく，様々な形態が誕生し，現在もその多様化は進んでいる．例えば，健康回復や維持，増進につながる観光はヘルスツーリズム，スポーツを体験または観戦することを目的とした観光はスポーツツーリズムと呼ばれる．旅行ブログエントリに対して，このような観光の形態の自動分類が実現すれば，世界各地の観光地でどのような形態の観光が可能か調べることができる．また，特定の形態に基づいた観光地の推薦や旅行計画も可能になると考えられる．そこで本研究では，6 種類の観光の形態を定義し，機械学習を用いて旅行ブログエントリをこれらの観光の形態に自動分類する手法を提案する．また，観光地ごとの特徴を明らかにするため，この分類結果を可視化するシステムを構築する．

2　観光の形態に基づいた旅行ブログエントリの自動分類

テキスト情報を用いて旅行ブログエントリなどの SNS を様々な観点から分類する手法がこれまでにも提案されている[1, 2]．しかし，ブログエントリ中に含まれる画像が分類の際に重要な情報になる場合も少なくない．例えば，「スキーをしたかったけどできなかった」という記述には「スキー」という言葉が含まれているが，ブログ著者は実際にスキーをしているわけではない．これに対し，スキーをしている画像があればブログ著者は確実にスキーをしていると判断できる．このように，ブログエントリを分類する際には，テキスト情報に加えて画像情報が分類の際の重要な判断材料になると考えられる．

人手で分類を行う際に根拠となる情報源は，上記で述べたテキストや画像のどちらかとなるが，旅行ブログエントリの内容によっては，テキストや画像から読み取ることができたものに関する外部知識が必要になることがある．例えば，姫路城を対象とした観光の場合，世界遺産を対象とした観光である「ヘリテージツーリズム」に分類できるが，ブログエントリ中のテキストや画像には，姫路城が世界遺産であるという情報が書かれていないことがある．これに対し，Wikipedia のような外部知識を与えることで，姫路城が世界遺産であるという情報が得られ，より精度の高い分類が期待できる．そこで，本研究では，テキスト中の単語集合に対し，Wikification[3, 4]を行い，テキスト情報と画像情報に加えてこの Wikification の結果を解析した結果を考慮した分類を行う．なお，Wikification とは，テキスト中の単語と Wikipedia エンティティをリンク付けすることである．

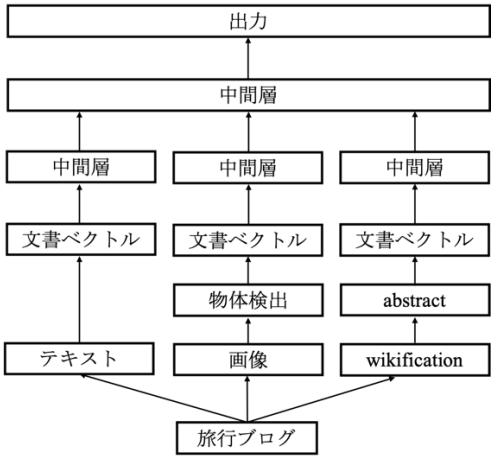

図 1: 分類機の概略図

本研究では，複数の入力データを考慮するため，それぞれの入力データに対して処理を行い，その処理結果を中間層で統合する分類器を構築した．分類器の概略図を図 1 に示す．この分類器では，まず，分類対象のブログエントリに含まれる画像に対し，画像認識技術を用いて物体検出を行う．次に，テキスト中の単語に対し，Wikification を行うことで，テキスト中の単語にリンク付けされた Wikipedia エンティティの該当ページの abstract（最初の段落）を抽出する．その後，テキストに含まれる単語集合，物体検出の結果から得られた単語集合および Wikipedia の abstract から得られた単語集合のそれぞれの入力データから，文書の分散表現を作成し，これらを入力としたニューラルネットワークで分類器を構築する．文書の分散表現を作成する手法については，SCDV(Sparse Composite Document Vector)[5]を用いた．SCDV では，単語ベクトルを GMM と IDF 値を考慮した新たなベクトルを生成し，生成した単語ベクトルの平均を文書ベクトルとする．SCDV の概略図を図 2 に示す．この図では，GMM を用いて各クラスタ a-e へクラスタリングを行い，その結果と IDF 値を考慮した単語ベクトルを平均することで文書ベクトルを生成している．

3　実験

提案手法の有効性を調べるため，実験を行った．

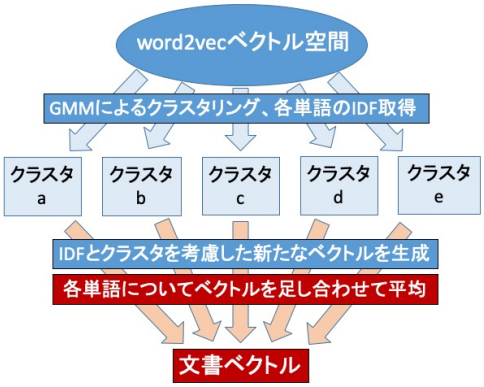

図 2: SCDV の概略図

データ

TravelBlog(travelblog.org)に存在する 2,017 件のブログエントリから，テキストと画像データを収集した．これらのデータに対し，6 種類の観光の形態に人手で分類したものを訓練および評価に用いる．

実験条件

単語の分散表現には，Google News を対象に Word2Vec を用いて獲得された 300 次元のモデルを用いる．画像認識は，Google Cloud Vision API[1] を用いる．Wikification には，Google Cloud Natural Language API[2] を用いる．また，ensemble 学習を用いて複数の分類器の出力結果を考慮した分類器を構築する．ベースライン手法として SVM を用いる．SVM に用いるカーネル関数は，テキスト，画像どちらも RBF カーネルとする．評価尺度には，micro 平均の精度・再現率・F 値を用いる．

実験結果

テキスト，画像および Wikipedia の abstract の組み合わせによる分類実験の結果を表 1 に示す．表 1 より，再現率と F 値では，テキストを用いた SVM が最も高い値となったものの，精度では ensemble(提案手法)で最も高い値 0.807 を得た．

表 1: 分類実験の結果

分類手法	精度	再現率	F 値
Ensemble(提案手法) ● SCDV(txt+img+wiki(abst)) ● SVM(txt) ● SVM(img)	**0.807**	0.179	0.293
Ensemble ● SCDV(txt) ● SVM(txt) ● SVM(img)	0.747	0.216	0.335
SCDV(txt+img+wiki(abst)) (提案手法)	0.752	0.218	0.338
SCDV(txt+img)(提案手法)	0.729	0.227	0.347
SCDV(txt)	0.639	0.169	0.268
SVM(txt)	0.654	**0.272**	**0.385**
SCDV(img)	0.725	0.140	0.235
SVM(img)	0.788	0.170	0.279
SCDV(wiki(abst))	0.528	0.116	0.191

4　可視化

本研究では，分類実験で使用したデータを学習に用い，TravelBlog のエントリ約 200,000 件に対して分類を行った．分類器には SVM(img)を用い，その結果を Google Earth にマッピングすることで実現した．実際にマッピングした結果を図 3 に示す．これは，奈良県の東大寺周辺を切り取ったものであるが，ヘリテージツーリズムを表す赤色の「建物」のマークが多く見られることから，この地域ではヘリテージツーリズムが多いことが確認できる．このように可視化することで，観光地の特徴や隠れた魅力の発見などが期待できる．

図 3: 奈良県の東大寺周辺の可視化画面

5　まとめ

本研究では，旅行ブログエントリ中のテキストと画像，Wikification による Wikipedia の情報を考慮し，定義した 6 種類の観光の形態に自動分類する手法を提案した．実験の結果，再現率と F 値では，ベースライン手法に劣るものの，ensemble 学習を用いた提案手法で，最も高い精度が得られた．また，実際に大量の旅行ブログエントリに対し分類を行い，可視化を行った．

参考文献

[1] K. Takahashi, *et al.*, “Analyzing Travel Behavior Using Multi-label Classification From Twitter,” Proc. of The 9th International Conference on Management of Digital EcoSystems, MEDES’17, 2017.

[2] K. Fujii, *et al.*, “Travellers’ behaviour analysis based on automatically identified attributes from travel blog entries,” Proc. of Workshop of Artificial Intelligence for Tourism, PRICAI, 2016.

[3] R. Mihalcea, *et al.*, “Wikify! Linking Documents to Encyclopedic Knowledge,” Proc. of The ACM Conference on Information and Knowledge Management, pp. 233–242, 2007.

[4] M. Yugo, *et al.*, “Wikification for Scriptio Continua,” Proc. of The 10th Edition of the Language Resources and Evaluation Conference, LREC, pp. 1346–1351, 2016.

[5] D. Mekala, *et al.*, “SCDV: Sparse Composite Document Vectors using soft clustering over distributional representations,” Proc. of Conference on Empirical Methods in Natural Language Processing, EMNLP, pp. 659-669, 2017.

[1] https://cloud.google.com/vision/?hl=en

[2] https://cloud.google.com/natural-language/?hl=en