SCDV を用いた観光地の口コミの可視化

2016SC002 安藤圭祐 2016SC078 清水涼太郎
指導教員：河野浩之

1　はじめに

近年, 娯楽として世界各地での観光が増えており, 国連世界観光機関 (UNWTO)[5] によると, 2017 年の国際観光客到着数の総数は 13 億 2,600 万人を記録し, 国際観光収入の総額は 1 兆 3400 億米ドルを記録し, 国際観光客到達数と国際観光収入はこれからも増加するであろうと考えられている．世界各地の観光地は文化や習慣の違いからさまざまな特徴があり, 自分の嗜好に合った観光地を見つけ出すのは困難である．こうした背景から観光に関する研究は需要が高く, レコメンデーションなどさまざまな観光に関する研究が盛んに行われている．

本研究では, 観光地の口コミの可視化を提案する．近年, Web 上で公開されている観光地の口コミサイトを利用して旅行を計画する人が増えているが, 膨大の量の口コミから観光地の特徴を得るのは困難となっている． そこで私たちは観光地の口コミの特徴を考慮できるように文書ベクトルを作成し, 観光地の口コミの可視化を行う．その後ベクトル化手法の評価をし，どの程度観光地の口コミが分類できているのかを検証する．

本研究は全 5 章で構成されており，2 章では文書ベクトル表現の関連研究，3 章では観光地の口コミの可視化の提案，4 章では観光地の口コミの可視化の評価，5 章ではまとめを示す．

2　文書ベクトル表現の関連研究

納村らの研究 [3] は Doc2Vec を用いて TF-IDF の問題を解決できることを検証した．実験の結果，Doc2Vec を用いることで同じタグの密集度合いを示す recall は，ランダムにクラスタに分類し得られた結果よりも高い値となり，関連の記事がまとまっているなどが見られた．しかし TF-IDF の問題を解決できるかの実験となっており，タグの密集度のみの検証となっている．

柴田らの研究 [4] は旅行ブログエントリ中のテキストおよび画像情報を用いて 6 種類の観光の形態に分類する手法を提案した．またエンティティリンキング技術を用いて Wikipedia エントリの情報をリンクできるようにした．実験の結果，SCDV を使用した結果が一番精度が高いものとなっている．

安藤らの研究 [1] はインクジェット関連特許を様々な手法で文書ベクトルを作成した．また機械学習を用いて様々な手法で文書分類を行った．実験の結果，SCDV による文書ベクトルを用いて XGBoost による文書分類が一番精度が高いものとなっている．

柴田らの研究 [5] と安藤らの研究 [1] では SCDV の精度が最も高くなっていた．そこで私たちは SCDV を用いて海外の観光地の口コミの可視化を提案する．その後 F 値を求め，観光地の口コミに SCDV を用いた場合でも良い精度が得られるか検証するため他のベクトル化手法とも比較を行う．

3　観光地の口コミの可視化の提案

本章では観光地の口コミの可視化の提案について記述する．3.1 節では観光地の口コミの可視化の概要について，3.2 節では学習に利用する観光地の口コミサイトについて記述する．

3.1　観光地の口コミの可視化の概要

本節では観光地の口コミの可視化の概要について記述する．今回の実験では様々な文書ベクトル表現を用いて観光地の口コミの可視化を行う．その後それぞれのベクトル化手法の評価をする．提案手法の概要を図 1 に示す．

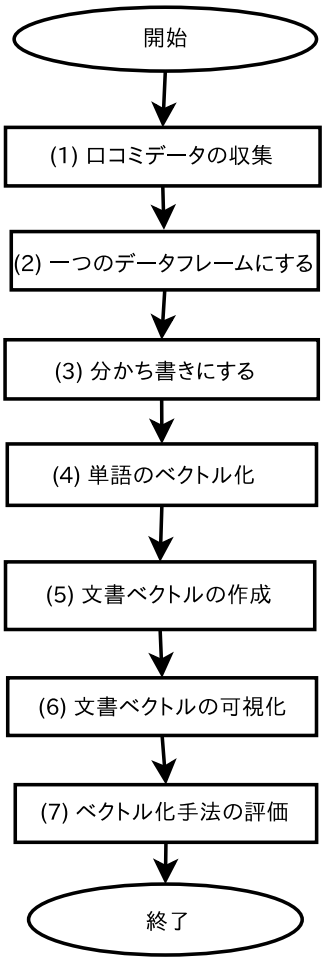

図 1　観光地の口コミの可視化の提案手法

(1) では 観光地の口コミサイトから学習するデータとなるテキストデータを収集する．口コミサイトから抽出す

ることで, 実際に旅行をした第三者の意見が反映した特徴ベクトルの獲得ができると考えられる．このときテキストデータはそれぞれ対応した国のディレクトリに分けて保存する．口コミデータを国ごとで分けることで, それぞれの国の特徴をわかるようにする．(2) では取得したデータを一つのデータフレームにする．(3) では単語のベクトル化を行えるようにテキストデータを分かち書きにする．英語では単語間にスペースがあることで機械的に単語を区別することが可能であるが, 日本語では単語間にスペースがないので機械的に単語の区別することが不可能であるため分かち書きにする．また数字や記号を省くことで単語のみのデータとし，より精度の高い学習をできるようにする．(4) では作成したデータフレームを基に単語のベクトル化を行う．(5) では文書ベクトルを作成する．これは各単語についてベクトルを足し合わせて平均することで文書ベクトルを得ることができる．(6) では観光地の口コミの可視化を行う．これにより各観光地の口コミの関係性を考察できるようにする．(7) では文書ベクトル化手法の評価を行う. これによりどの程度観光地の口コミが分類できているのかを考察できるようにする．

3.2 学習に利用する観光地の口コミ

本節では学習に利用する観光地の口コミについて記述する．これはそれぞれの単語に意味を持たせるために単語のベクトル化をし, その後文書ベクトルを作成するために利用する．今回の実験では 4travel を用いる．4travel, じゃらん net, トリップアドバイザーの 3 つの観光地の口コミサイトの比較を表 1 で示す.

じゃらん net は国内の観光地の口コミのみとなっているが 4travel とトリップアドバイザーは海外の観光地の口コミが多くなっている．また 4travel の口コミのマイル獲得条件が 1 件 100 文字以上であるのに対し, トリップアドバイザーは 1 件 50 文字以上であるので, 4travel の方が比較的口コミの文字数が多く, 観光地の特徴を多く含んだ文書ベクトルを作成できると考えられる．そのため, 今回は 4travel のテキストデータを基にコーパスを作成する.

表 1 観光地の口コミサイトの比較

口コミサイト	特徴
4travel	海外の観光地の口コミが多い 口コミの文字数が多い
じゃらん net	国内の観光地の口コミのみである
トリップアドバイザー	海外の観光地の口コミが多い 口コミの文字数が少ない

4 観光地の口コミの可視化の評価

本章では観光地の口コミの可視化の評価について記述する．4.1 節ではデータの取得方法について，4.2 節では Word2Vec による学習について，4.3 節では SCDV の作成について，4.4 節ではデータの可視化について，4.5 節ではベクトル化手法の評価について記述する．

4.1 データの取得方法

本節ではデータの取得方法について記述する．本実験において，4travel の口コミをスクレイピングし，8 つの国の口コミの可視化をする．またイギリス，フランス，ハワイ，香港，韓国，シンガポール，台湾，タイの観光地の口コミを取得し，一つの国につき観光スポットランキング上位 5 つの観光地の口コミを一つの観光地につき 50 件の口コミを取得し，8 つの国の口コミの合計 2000 件のデータのスクレイピングを行う．スクレイピングは Python3.8.0 でスクレイピングに特化した機能を持った Python のライブラリである Beautiful Soup を利用する．

一つのページにつき 10 件の口コミが書かれており，10 件の口コミが書かれているテキストファイルを取得することができる．このテキストファイルを一件の口コミが書かれたテキストファイルにするためにファイル分割をするソフトウェアである div8 を利用する．これにより 10 件の口コミが書かれているテキストファイルを 1 件ごとのテキストファイルに分解することができる．分解したテキストファイルはそれぞれ対応したディレクトリに入れる．本実験では「British」「France」「Hawaii」「HongKong」「Korea」「Singapore」「Taiwan」「Thailand」の 8 個のディレクトリを作成する．「British」のディレクトリには「大英博物館，ナショナルギャラリー，バッキンガム宮殿，タワーブリッジ，ビッグベン」の口コミ，「France」のディレクトリには「ルーヴル美術館，エッフェル塔，凱旋門，ノートルダム大聖堂，オルセー美術館」の口コミ，「Hawaii」のディレクトリには「ダイヤモンドヘッド，ワイキキビーチ，カイルアビーチパーク，ラニカイビーチ，ハナウマ湾」の口コミ，「HongKong」のディレクトリには「ビクトリアピーク，香港ディズニーランド，シンフォニーオブライツ，ネイザンロード，ビクトリア湾」の口コミ，「Korea」のディレクトリには「景福宮，N ソウルタワー，北村韓屋村，甘川洞文化村，昌徳宮」の口コミ，「Singapore」のディレクトリには「マーライオン公園，ガーデンズバイザベイ，ナイトサファリ，チャイナタウン，マリーナベイサンズライト＆ウォーターショー」の口コミ，「Taiwan」のディレクトリには「九分の町，龍山寺，国立故宮博物院，国立中正紀念堂，台北 101 展望台」の口コミ，「Thailand」のディレクトリには「ワットポー，ワットアルン，ワットプラケオ，チャオプラヤー川，ワットパークナム」の口コミを入れる．

4.2 Word2Vec による学習

まず 4.1 節で作成した 8 個のフォルダに入っているテキストファイルを一つのリストにまとめ,「pandas」を用いて一つのデータフレームを作成する．次に作成したデータフレームを MeCab を用いて分かち書きにする．MeCab はインポートすることで利用することができ，モードを

「Owakati」とし，分かち書きのみを出力されるようにする．また数字や記号はなくすように設定する．これらの作業により無駄な文字をなくし空白で単語ごとに分解することができる．その後作成した分かち書き文を用いて Word2Vec で単語のベクトル化を行う．今回の実験では Word2Vec のパラメータを次元数を 200，学習に使う前後の単語数を 10 と設定する．また 5 回未満登場する単語の破棄を設定したため 2711 単語で学習を行う．

4.3 SCDV の作成

本節では SCDV の作成について記述する．SCDV とは Word2Vec のベクトル空間をクラスタリングし，各単語がどのトピックに属しているのかを考慮した分散表現である．SCDV の作成手順を以下に示す．

(1) Word2Vec ベクトル空間を取得
(2) GMM によるクラスタリング
(3) 各単語の IDF 値を取得
(4) IDF 値とクラスタを考慮した新たなベクトルを生成
(5) 各単語についてベクトルを足し合わせて平均し文書ベクトルを生成
(6) 文書ベクトルをスパースする

まず (1) では前節で作成した Word2Vec のモデルを取得し，(2) では単語ベクトルを「GaussianMixuture」を利用し混合ガウスモデル (GMM) でクラスタリングする．GMM を使うことで，データセットをクラスタごとに分けられるだけでなく，データセットの確率密集分布を得ることができる．これにより各単語のクラスタと各単語の各クラスタに属する確率を求めることができる．次に (3) では「TfidfVectorizer」を利用し IDF 値を求める．「TfidfVectorizer」とは「CountVectorizer」と「TfidfTransformer」の機能も持つものである．「CountVectorizer」はトークン化とカウントができるメソッドであり，「TfidfTransformer」は正規化ができるメソッドである．その後 (4) では各単語ベクトルと各単語の各クラスタに属する確率，TF-IDF 値を掛け合わせ確率で重み付けした Word2Vec を求める．そして (5) では文章の構成単語について平均を求め文書ベクトルを作成する．最後に (6) では文書ベクトルをスパースする．スパースとは「すかすか」,「少ない」を意味し，今回の実験では学習した単語数が少ないため行う．この方法はデータ量に比べて大量の学習パラメータを用意し，パラメータの自動抽出を行うことによって，単純で過学習を起こさないモデルを得る方法である．これにより SCDV を得ることができる．

4.4 文書ベクトルの可視化

本節では文書ベクトルについて記述する．可視化を行うためにまずは高次元データの次元を圧縮するために t-SNE を用いる．その後，Python におけるグラフ描写の標準的なライブラリである matplotlib を用いてデータの描写を行う．Word2Vec による可視化の結果を図 2，SCDV による可視化の結果を図 3 に示す．散布図の軸は文書ベクトルの数値を示しており，クラスタの青色が British，橙色が France，緑色が Hawaii，赤色が Hongkong，紫色が Korea，茶色が Singapore，桃色が Taiwan，鼠色が Thailand を示している．

図 2 ではフランスのクラスタとシンガポールのクラスタ，台湾のクラスタはクラスタ同士が重なっており観光地の関係性を考察することが困難である．図 3 では，クラスタ同士の重なりがなくなっており，各クラスタがいくつかのまとまりごとに分布していることがわかる．この結果からイギリスのクラスタとフランスのクラスタが近くに分布していることから 2 つの国で口コミが類似していることが読み取ることができる．またシンガポールのクラスタとハワイのクラスタ，韓国のクラスタは一箇所にまとまって分布していることからそれぞれのクラス内で観光地の口コミが類似していることが読み取ることができる．

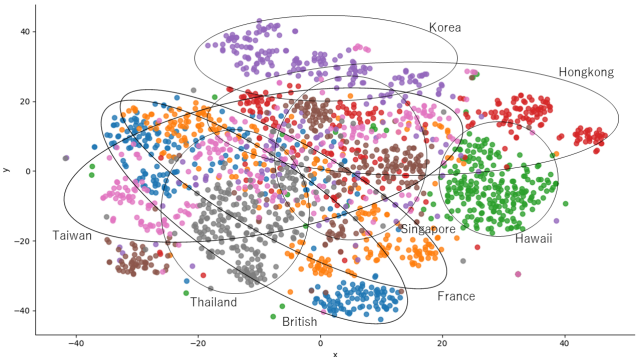

図 2 Word2Vec による口コミの可視化

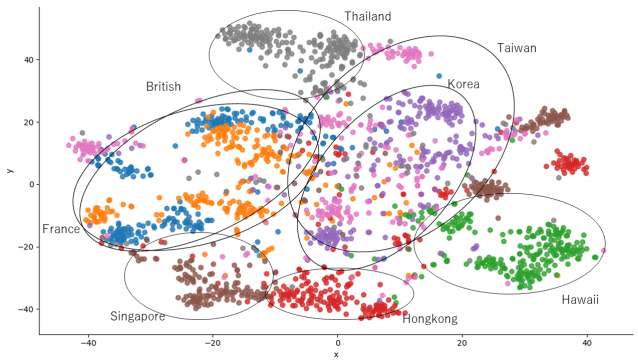

図 3 SCDV による口コミの可視化

次に文書ベクトルの口コミの内容の視点から考察する．図 3 の左側でイギリスと台湾のクラスタが隣り合って分布していることがわかる．このクラスタがどの観光地の口コミを示しているのかを知るためにイギリス内で観光地ごとにクラスタの色を分けてどの観光地の口コミを表しているのかをわかるようにし，台湾でも同様に行う．その結果，

イギリスの観光地はバッキンガム宮殿であり，台湾の観光地は国立中正紀念堂を表していた．このとき 2 つの観光地の口コミの内容は共通して衛兵の交代式に関する口コミが多く見られた．この結果から 2 つのクラスタが近くに分布していたことが考えられる．

4.5 ベクトル化手法の評価

本節ではベクトル化手法の評価について記述する．あらかじめ「train-test-split」を用いてデータセットをトレーニング用データとテスト用データに分割しそれぞれで文書ベクトルを作成する．このときテストデータを 30 %と設定する．これはトレーニング用データと同じデータでテストを行ってしまうと適切なスコアを求めることができなくなるためである．今回の実験では LightGBM と classification-report を利用し F 値を求めベクトル化手法の評価をする．LightGBM とは決定木アルゴリズムに基づいた勾配ブースティングの機械学習データフレームである．classification-report とは precision(適合率), recall(再現率)，F 値を求められるパッケージである．適合率とはクラスと判断したすべてのデータのうち，実際にそのクラスであった割合であり，再現率とはあるクラスのすべてのデータに対して実際にそのクラスであると判断できる割合であり，F 値とは再現率と適合率の調和平均を取った値であり数値が高い程分類の精度が高いと言える．Word2Vec の評価を表 2，SCDV の評価を表 3 に示す．

表 2, 表 3 から Word2Vec での F 値の平均は 0.817010 となっており, SCDV での F 値の平均は 0.897347 となっており，F 値の平均は約 8 %上昇していた．この結果から SCDV を用いた方が分類の精度が高いことがわかる．またフランスの F 値は約 14 %上昇し，韓国の F 値は約 13 %上昇し，シンガポールの F 値が約 12 %上昇しており他の国と比べて SCDV の影響を大きく受けていることがわかる．しかしハワイの F 値は約 2 %の上昇で香港の F 値の向上は見られなかった．これは，ハワイの Word2Vec での F 値が 0.913580 となっており，香港の Word2Vec での F 値が 0.869565 となっており元々の Word2Vec での F 値が高かったためであると考えられる．

表 2 Word2Vec の評価

	precision	recall	F 値
British	0.773333	0.852941	0.811189
France	0.816901	0.734177	0.773333
Hawaii	0.891566	0.936709	0.913580
Hongkong	0.857143	0.882353	0.869565
Korea	0.790123	0.820513	0.805031
Singapore	0.818182	0.777778	0.797468
Taiwan	0.740260	0.730769	0.735484
Thailand	0.859375	0.820896	0.839695
weighted avg	0.817830	0.817726	0.817010

表 3 SCDV の評価

	precision	recall	F 値
British	0.861111	0.911765	0.885714
France	0.911392	0.911392	0.911392
Hawaii	0.925926	0.949367	0.937500
Hongkong	0.890625	0.838235	0.863636
Korea	0.913580	0.948718	0.930818
Singapore	0.935897	0.901235	0.918239
Taiwan	0.847222	0.782051	0.813333
Thailand	0.887324	0.940299	0.913043
weighted avg	0.897771	0.897993	0.897347

5 まとめ

本研究では観光地の口コミデータを用いて Word2Vec で単語のベクトル化を行い，文書ベクトルを作成した．また Word2Vec のモデルを用いて SCDV を取得し文書ベクトルを作成し，それぞれの文書ベクトルの可視化を行った．またそれぞれのベクトル化手法の評価を F 値で比較した．

実験の結果，Word2Vec で作成した文書ベクトルの可視化ではクラスタ同士が重なっていたが，SCDV で作成した文書ベクトルの可視化ではクラスタ同士の重なりがなくなっており，各クラスタがいくつかのまとまりごとに分布していた．また Word2Vec での F 値の平均は 0.817010 となっており, SCDV での F 値の平均は 0.897347 となっており，F 値の平均は約 8 %上昇していた．これらの結果から SCDV を用いた場合の方が観光地の口コミの可視化に適していることが検証できた．

参考文献

[1] 安藤俊幸, 桐山 勉：“分散表現学習を利用した効率的な特許調査 文書のベクトル化方法と文書分類への応用,” 第 16 回情報プロフェッショナルシンポジウム, 情報科学技術協会, pp. 31-36, 2019.

[2] 森巧尚: “Python2 年生 スクレイピングのしくみ 体験してわかる！会話で学べる！,” 株式会社 翔泳社 (2019).

[3] 納村聡仁, 沼尾正行, 福井健一：“語順を基にした分散的意味表現による観光文書表現の検証,” 2016 年度人工知能学会全国大会 (第 30 回), pp. 1-4, 2016.

[4] 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分類と可視化,” 第 135 回 IFAT 研究発表会, 情報処理学会, pp. 1-8, 2016.

[5] UNWTO: UNWTO Tourism Highlights 2018 Edition 日本語版, https://unwto-ap.org/wp-content/uploads/2019/01/Tourism-HL-2018.pdf, 参照 Jan 4, 2020.