観光レビューからの観点抽出の検討

Examination of Aspect Extraction from Tourism Reviews

小林らんう
Ranu Kobayashi

岡山大学 太田研究室
Ohta laboratory, Okayama University

概要 本研究では観点抽出のためにじゃらん net の観光レビューから，エンティティを抽出し，そのエンティティをクラスタリングする．クラスタリング後に各クラスタのエンティティを確認し分析する．さらに，ラベル付けが可能であればクラスタの要素に見合うラベルを付与する．ラベル付けができたクラスタはそのラベルが観点としてふさわしいかどうかを考察する．

1　はじめに

近年インターネットの普及に伴い，じゃらん net[1]や楽天トラベル[2]などの観光サイトに観光レビューを残す観光客が多くなっている．また，観光レビュー以外にも Twitter や Instagram などの SNS に観光の感想などを書き込む人も多い．観光スポットや観光ルート推薦の研究ではこれらのレビューやツイート(SNS 上の書き込み)などの書き込みといくつかの観点を用いて，観光スポットを分析する手法がしばしば用いられる．

例えば，野本ら[1]は，「食事」，「景観」，「購買」，「体験」，「設備」，「混雑」，「交通」の計 7 つの観点を定義し，レビューと観点を比較することでスポットの特徴をとらえた．角上ら[2]は，「食事」，「土産」，「景観」，「行動」の計 4 つのカテゴリ(観点)を定義し，収集したツイートをそのカテゴリに分類し，分類済みのツイートを用いてスポットのカテゴリスコアを算出した．このカテゴリスコアは当該研究で推薦スポットを評価する指標の内一つである．杉浦ら[3]は，被験者に観光スポットに関する質問をし，それに対する被験者の回答(評価表現)を収集した．収集された評価表現から「世界遺産」や「落ち着ける」などの評価要因(観点)を 137 得て，さらにそれらを「気分」「体験」「雰囲気」「スポットの特徴」の 4 つのカテゴリに分類した．次に別の被験者に推薦対象の 150 箇所の京都の観光スポットのうち，行ったことがあり好ましいスポットを入力させ，それらのスポットについて 137 の評価要因に関する質問に 7 段階(1:全く当てはまらない~7:非常によく当てはまる)で回答させた．質問は「神社・仏閣である」のようなものである．またモデルを使用するユーザは最初にカテゴリを選び，次に評価要因を 1 つ以上選ぶ．これらのアンケートの結果とユーザの入力を用いてベイズ定理に基づく式でスコアを算出し，ユーザの入力に対応するランキングを表示する推薦システムを構築した．篠田ら[4]は旅行ブログを「買う」，「食べる」，「体験する」，「見る」，「泊まる」の合計 5 種類の観光タイプ(観点)に自動分類し可視化を行った．

しかし，観点を人手で決める場合，偏りが生じる可能性が高い．例えば，「混雑」，「立地」，「食事」，「景観」のように 4 つの観点を定めても，「伊勢神宮に行って見つけました 養蜂体験！！」のようなツイートはそのいずれにも分類できない．このツイートは，例えば野本ら[1]が定義した「体験」のような観点があれば分類可能になる．このように観点を適切に定義することは後の利用に大きな影響を与える．本研究では，レビュー中から観点の候補となる単語を自動抽出し，多様な観点の抽出を目指す．

2　レビューからの観点抽出

本研究では観点抽出のために観光サイトのレビューからエンティティを抽出し，そのエンティティ群を k-means 法によりクラスタリングする．ここでエンティティとは文章中に含まれる人物名，ランドマークなどの固有名詞や，レストラン，競技場などの普通名詞のことである．また，クラスタリング後に，クラスタの要素であるエンティティを分析しそのクラスタに見合うラベルを著者が人手で付与する．そして，ラベル付けされたクラスタにおいてそのラベルが観点としてふさわしいかどうかを考察する．以下の 2.1 節で分析する観光レビュー，2.2 節でエンティティ抽出，2.3 節でエンティティのクラスタリングについて詳しく述べる．

2.1　分析する観光レビュー

本研究では Web スクレイピングを用いてじゃらん net から観光レビューを取得する．本稿では 2022 年 7 月 5 日時点の岡山の「7 月にオススメランキング」[3]のうち上位 30 ヵ所のレビューを収集した．表 1 に「岡山後楽園」のレビュー

表 1:岡山後楽園のレビューの例

レビュー1	三大名園の一つです。とてもきれいな風景で、いやされました。散策がとても心地よく、また行きたいと思いました。
レビュー2	家族連れで行きました。広大な庭園は見応えがあり庭園から岡山城も見えました。お茶屋もありそこでぜんざい&ほうじ茶のセットを食べました。庭園&城を眺めながらゆっくりお茶が出来良い時間でした。機会があればもう一度訪れたい場所です。

1 https://www.jalan.net/

2 https://travel.rakuten.co.jp/

3 https://www.jalan.net/kankou/330000/page_1/

表 2:レビュー1 から抽出されたエンティティ

エンティティ名(name)	三大名園	風景	散策	三	一
タイプ(type)	OTHER	OTHER	OTHER	NUMBER	NUMBER
重要度(salience)	0.145	0.143	0.143	0.101	0.099

表 3:レビュー2 から抽出されたエンティティ

エンティティ名(name)	家族連れ	庭園	見応え	岡山城	茶屋	ぜんざい	
タイプ(type)	PERSON	LOCATION	OTHER	LOCATION	LOCATION	CONSUMER_GOOD	
重要度(salience)	0.145	0.143	0.101	0.099	0.051	0.048	
エンティティ名(name)	ほうじ茶	時間	セット	茶	機会	庭園&城	場所
タイプ(type)	OTHER	OTHER	OTHER	OTHER	OTHER	OTHER	LOCATION
重要度(salience)	0.048	0.040	0.040	0.039	0.038	0.035	0.030

を例として示す．収集したレビューは計 19307 件となった．

2.2 エンティティ抽出

エンティティ抽出のために Google Cloud の Natural Language API[4]を用いる．これを用いることでエンティティの名前のほかに，エンティティのタイプ(type)と，重要度(salience)を抽出できる．Type は PERSON，LOCATION などの 13 項目からなり，重要度は[0.0, 1.0]の値をとり， 1.0 に近いほど重要なエンティティであることを示す．本手法で抽出されたエンティティは合計で重複を含めて 169,105 となった．本稿ではこれらのエンティティから重複を除き，さらに重要度が 0.5 以上のエンティティ 473 をクラスタリング対象として選んだ．抽出例として表 2 に表 1 の観光レビュー1 の，表 3 に表 1 の観光レビュー2 のエンティティを示す．

2.3 クラスタリング

2.2 節で説明したエンティティをクラスタリングするためにまずエンティティを分散表現に変換する．そのために，fastText[5]を用いる．fastText は Word2vec を基盤に作られている自然言語処理ライブラリで，単語を分散表現に変換できる．本研究では Wikipedia[6]と Common Crawl[7]で事前学習された fastText の日本語モデル[8]を用いる．

エンティティの分散表現を得るには，そのエンティティが語彙として学習済みモデルに含まれなければならない．そのためもし語彙にそのエンティティが存在しなければ本研究ではそのエンティティを Juman++[9]を用いて形態素解析し，ベクトル演算を用いて分散表現を求める．例えば，「家族連れ」は形態素解析で「家族」と「連れ」に分けられるため，「家族」と「連れ」の分散表現をそれぞれ求め足し合わせて「家族連れ」の分散表現とする．この手法を用いた結果，エンティティ 473 のうち 450 の分散表現を獲得することができた．分散表現を得ることができた 450 のエンティティの内 285 はエンティティそのものが語彙に存在し，165 はベクトル演算を用いて分散表現を得た．分散表現を得ることができなかった 23 のエンティティは英語のエンティティや誤字のエンティティなどであった．例えば「しょぼい」の誤字と思われる「しょぽい」は分散表現を得ることができなかった．

次にエンティティの分散表現を k-means 法を用いてクラスタリングする．エンティティが各クラスタに均等に分類されると仮定したときクラスタ毎の要素数が約 12 になるようにクラスタ数を設定したところクラスタ数は 40 となった．fastText やクラスタリングは Google Colaboratoy 上の Python3.7.13 を用いて実行した．

3 エンティティ分類の結果と考察

観光レビューから抽出したエンティティのクラスタリング結果のうち，著者がラベル付け可能であったクラスタを表 4 に，ラベル付けが不可能であったクラスタを表 5 に示す．ただし，要素数が 1 の 12 クラスタは省略する．

ラベルを付けでは，クラスタの要素中最も出現頻度が高い単語をラベルとして付与する．例えば，表 4 のラベル「瀬戸」のクラスタはすべてのエンティティに「瀬戸」という単語が含まれているためラベルは「瀬戸」とした．表 4 のラベル「岡山」のクラスタは「岡山」が含まれるエンティティが 3 つ，「倉敷」が含まれるエンティティが 2 つでラベルを「岡山」とした．一方表 4 のラベル「倉敷」のクラスタは「岡山」が含まれるエンティティが 2 つ，「倉敷」が含まれるエンティティが 6 つでラベルを「倉敷」とした．例外として，表 4 中のラベル「名園」は単語に「三大」が共通しているが，「三大」ではラベルの要素を説明できないためラベルは「名園」とした．また，表 4 中のラベル「感覚」や「尺度」のように，要素にはそのラベルの単語は含まれないラベルも存在する．

表 5 はラベル付けが困難なクラスタを示しており，これらのうち要素数が 30 を超えるクラスタ番号 1 から 4 までの

4 https://cloud.google.com/natural-language?hl=ja

5 https://fasttext.cc/

6 https://www.wikipedia.org/

7 https://commoncrawl.org/

8 https://github.com/facebookresearch/fastText/blob/master/docs/crawl-vectors.md

9 https://nlp.ist.i.kyoto-u.ac.jp/?JUMAN%2B%2B

表 4:ラベル付けできたクラスタ(23)

クラスタラベル (要素数)	エンティティ
瀬戸(3)	瀬戸大橋，瀬戸内海，瀬戸
海(6)	海沿い，渋川海岸，海岸道路，国立公園，海岸
子供(12)	来館者，子供達，子どもたち，魚たち，子供連れ，女子たち，子供，こどもたち，友達親子，教科書，子供たち，家族連れ
岡山(7)	岡山駅，岡山城，JR 倉敷駅，倉敷駅，岡山市，香川県，吉備津神社
温泉(7)	温泉，温泉地，天然温泉，駐車場，湯原温泉，温泉場，海水浴場
名園(3)	日本三大名園，三大名園，日本三大庭園
入場料(6)	入場料，入場料金，入場無料のんびり，入館料，入場無料，入園料
倉敷(10)	岡山，岡山後楽園，倉敷紡績，倉敷，アリオ倉敷，倉敷ラーメン，倉敷観光，倉敷インター，高松，白山神社
感覚(5)	ゴチャゴチャ感，解放感，開放感，清潔感，開放目的
旅行(4)	日帰りバス旅行，一泊旅行，旅行，家族旅行
尺度(3)	広さ，大きさ，高さ
安全(2)	家内安全，安全祈願
高原(2)	高原気分，高原
庭園(3)	幻想庭園，庭園，日本庭園
屋(3)	お土産屋，抹茶屋，食べ物屋
動物(5)	動物たち，動物園，オリーブ園，日本三名園，動物
観光(6)	船観光船，観光客，観光地，観光船，観光スポット，観光
蒜山(3)	蒜山，蒜山ジャージーソフトクリーム，蒜山高原
犬(3)	犬，猫カフェ柴犬カフェふれあい，柴犬
絶叫(3)	絶叫マシン，絶叫，絶叫マシーン
美術館(5)	大原美術館，所蔵品，美術館，食料品，美術品
美観地区(2)	美観地区，倉敷美観地区
休憩(4)	途中休憩，休憩スペース，休憩場所，訪問場所

クラスタに関してはその要素をもう一度 k-means 法でクラスタリングする．各クラスタ数はそれぞれ 15，7，6，3 とした．そのクラスタリングした結果を表6から表8に示す．ただし，クラスタ番号 4 のクラスタはもう一度クラスタリングを行ったところ，1 つのクラスタを除く他のクラスタの要素数がすべて 1 となり，クラスタリング前とあまり変わらない結果となった省略する．また表 6，表 7，表 8 にはラベル付けが可能であったクラスタのみを示す．表 6，表 7，表 8 の結果から，ラベル付け可能なクラスタリングをもう一度行うことにより，まとまりのあるクラスタが新たに得られたことがわかる．表 9 に表 4，6，7，8 のクラスタラベルをまとめる．これらのラベルが観点としてふさわしいかどうかを分析する．「瀬戸」は単に地域を指しているだけであり，観光スポット推薦においての観点としてはふさわしくないと艦会える．同様の理由で「岡山」，「倉敷」，「蒜山」，「美観地区」，「四国」もふさわしくない．また「温泉」，「美術館」，「庭園」，「社寺」はスポットの種類を表し，観光客の行動に影響を与えるため観点としてふさわしいと考える．他にも「海」や「子供」のように付加的な情報を与えることができるラベルも観点としてふさわしい．逆に，「旅行」や「屋」のようにスポットの付加情報としてあいまいなラベルは観点としてふさわしくない．これらを整理して著者が観点としてふさわしいと判断したラベルを表 9 中に太字で示す．

表 5:ラベル付けできなかったクラスタ(5)

クラスタ番号	エンティティ
1(155)	テーマパーク，ショップ，ゴールデンウィーク，四国，ドイツ，エスカレータ，サンバ，愛媛，スニーカー，居心地，ホテル，…
2(76)	滝，路地裏，展望台，城，列車，道，雪恋まつり，散策，宿，山頂，庭，紅葉，面影，…
3(64)	昼間，娘，花見シーズン，小雨，夕暮れ時，出張，一生，バカ，友達，樹種積，雨，…
4(33)	多く，機会，形，スイーツ食べ放題，もの，場所，季節がら，感動，展示物，趣，印象，目的，時期，眺め，こと，隠れ，予想，…
5(4)	受胎告知，芝滑り，定期訪問，ゲームプラザおすすめ

エンティティのクラスタリングでは，まとまりのあるクラスタが生成される一方，エンティティが一部のクラスタに偏ることも確認された．また，クラスタの中には全くまとまりのないものもあった．今後は SVM や ward 法等，k-means 法以外のクラスタリング方法を試してみる必要があると考えている．また，クラスタのラベル付けは著者が行ったため，その妥当性を客観的に評価する必要がある．

表 6:クラスタ番号 1 のクラスタリング結果

クラスタラベル (要素数)	エンティティ
自然(20)	外，石山寺，八幡宮，山城，海，後楽園，水，花，鷲羽山，みどりゆたか，桜，梅，緑，夫，白鳥，孫，夏，さくら，モネ，真下
四国(3)	四国，愛媛，香川
雰囲気(13)	テーマパーク，パワースポット，気持ち，居心地，エリア，雑誌，ツアーガイド，他，雰囲気，パワー，名，サービスエリア，ボランティアガイド
飲食(12)	乗馬，牛，ドイツビール，手作りこんにゃく，野菜，チーズフォンデュ目当て，魚，カニ，福袋，フードコート，軽井沢バーガー，出店

表 7:クラスタ番号 2 のクラスタリング結果

クラスタラベル (要素数)	エンティティ
家(16)	宿，足湯，砂湯，住宅地，施設，煉瓦造り，駅，車，町，食堂，敷地，改装，店，建物，資料，農園
風景(20)	展望台，列車，芝生，散策，写真，紅葉，面影，町並み，景色，景観，地図，ぶらぶら散歩，かけ灯り，絵，街並み，絶景，絵画，散歩，風景，夜景
社寺(16)	滝，回廊，道，社殿，靴，登山，神殿，山頂，坂道，遊具，坂，階段，山道，鳥居，参拝，廻廊
路地(5)	路地裏，店内，商店街，館内，路地
歴史(17)	城，橋経，寺，庭，橋，公園，館，城マニア，雪，神社，歴史，狛犬，池，門，塔，王国，小さな八幡宮

表 8:クラスタ番号 3 のクラスタリング結果

クラスタラベル (要素数)	エンティティ
混雑 (15)	夕暮れ時，混雑，休日，殺伐，時間，空気，体力，人混み，人波，仕事前，空間，自然，退屈，息，足
気温(2)	気温，温度
雨(4)	小雨，雨，大雨，天気

表 9:クラスタラベルのまとめ

瀬戸	**海**	**子供**	岡山	**温泉**	**名園**	**入場料**	
倉敷	**感覚**	旅行	**尺度**	**安全**	**高原**	**名園**	屋
動物	観光	蒜山	**犬**	**絶叫**	**美術館**		**休憩**
美観地区		**自然**	四国	**雰囲気**		**飲食**	家
風景	**社寺**	**路地**	**歴史**	**混雑**	**気温**	**雨**	-

4 まとめ

本稿では，観光レビューから抽出したエンティティをクラスタリングすることで観光推薦等に利用可能な観点が抽出できるか検討した．岡山県の観光レビューから 473 のエンティティを抽出し，そのうち fastText で分散表現が得られた 450 のエンティティを 40 のクラスタに分類することができた．しかし 40 クラスタのうち 12 クラスタは要素数が 1 となり，さらに 5 クラスタはまとまりのないクラスタとなった．まとまりのない 5 クラスタの内 3 つを再度クラスタリングすることにより新たなクラスタを得ることができた．これらのクラスタに著者がラベル付けし，さらにそのうち観点となりうるものを選別した．

今後は著者のラベル付けの妥当性を客観的に評価するとともに，他のクラスタリング方法の利用も検討したい．さらに観点抽出の研究を進めて観光スポット推薦の研究に貢献したい．

参考文献

[1] 野本輝，上野史，太田学，観光レビュー文を用いた穴場スポットの発見，DEIM Forum 2022 B43-3，2022.

[2] 角上直哉，新妻弘崇，太田学，観光スポットの訪問目的を考慮した観光ルート推薦の一手法，DEIM2020 J1-1，2020.

[3] 杉浦孔明，岩橋直人，芳賀麻誉美，堀智織，階層型評価構造に基づく観光スポット推薦システムの構築と長期実証実験，観光情報学会第 8 回研究発表会講演論文集，pp.9-12，2013.

[4] 篠田広人，篠田有基，難波英嗣，石野亜耶，竹澤寿幸，旅行者の感情に基づいた観光スポット推薦，観光情報学会第 20 回研究発表会，pp.1-4，2019.